COMPUTATIONAL METHODS FOR GENOMIC VARIANT CALLING AND ANALYSIS

BY

DANIEL PAUL WICKLAND

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Informatics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2019

Urbana, Illinois

Doctoral Committee:

Professor Matthew E. Hudson, Chair
Associate Professor Yan W. Asmann, Mayo Clinic
Research Assistant Professor Liudmila S. Mainzer
Professor Stephen P. Moose
Professor Emerita Lila O. Vodkin

ABSTRACT

The development of short-read, next-generation sequencing (NGS) has revolutionized biological research, agriculture and medicine, enabling innovations such as genomic selection to raise crop yields and precision medicine to diagnose and treat disease. The genetic polymorphisms identified by this high-throughput sequencing can serve as markers for association with phenotypic traits. Variant calling refers to the process of detecting genetic polymorphisms based on analysis of genome sequence data output by NGS technology. The projects described here investigate these analysis methods.

Chapter One reviews variant calling and its application to human and plant genomic data. It opens by detailing the generation of sequence reads from biological samples and the conversion of those reads to meaningful data, emphasizing the importance of tool selection for analysis. Next, the use of sequencing to identify genetic risk factors in the context of Alzheimer's disease is reviewed. The chapter concludes by describing the application of sequencing to analysis of plant genomes.

Chapter Two presents a study of the impact of batch effect and study design on identification of genetic risk factors in human sequencing data. Sequencing-based searches for disease-associated variants require large sample sizes to achieve sufficient statistical power, but they often entail batch effects and biases from study design, both of which hinder the ability to detect true genotype-trait associations. We studied batch effects and confounding variables in whole-exome data from the Alzheimer's Disease Sequencing Project and demonstrated that both significantly impacted the association analysis. In particular, we identified variants with novel disease associations that may have been influenced by population stratification and a confounding effect of age.

Chapter Three reports a comparison of genotyping-by-sequencing (GBS) analysis methods on plant data. As a reduced-representation sequencing method to identify genetic variants and quickly genotype samples, GBS produces extensive missing data and requires complex bioinformatics analysis, particularly in the context of plants, which have highly variable

ploidy and repeat content. To address issues identified with existing methods, we developed GB-eaSy, a GBS bioinformatics pipeline that incorporates widely used genomics tools, parallelization and automation to increase the accuracy and accessibility of GBS data analysis. A comparison of five GBS pipelines using low-coverage sequence data from soybean demonstrated that GB-eaSy rapidly and accurately identified the greatest number of variants. In addition, the unexpectedly low convergence between the five analysis methods but generally high accuracy indicated that the workflows arrived at largely complementary sets of valid variant calls.

TABLE OF CONTENTS

CHAPTER 1: GENOMIC VARIANT CALLING WITH APPLICATION TO HUMAN AND PLANT DATA

OVERVIEW OF GENOMIC VARIANT CALLING

The development of short-read, next-generation sequencing (NGS) has revolutionized biological research, agriculture and medicine, enabling innovations such as genomic selection to raise crop yields and precision medicine to diagnose and treat disease. The genetic polymorphisms identified by this high-throughput sequencing can serve as markers for association with phenotypic traits. Variant calling refers to the process of detecting genetic polymorphisms based on analysis of sequence reads output by NGS technology.

NGS technology relies on sequencing by synthesis, in which short segments of cut DNA serve as templates for generation of complementary sequences composed of fluorescently tagged nucleotide bases (Nielsen et al. 2011). During synthesis of these complementary sequences by DNA polymerase, the addition of each tagged nucleotide is recorded by a camera. The sequencer outputs strings of nucleotide bases, or "reads," corresponding to these sequences, and each nucleotide base is assigned a quality score that measures the confidence of the base call. This high-throughput procedure occurs simultaneously for tens of millions of DNA templates in a single sequencing machine.

In order to determine the genomic origin of the sequence reads output by the sequencer, they must be assembled into larger contigs, often with the aid of a previously assembled and annotated reference genome. Alignment refers to the computational process of mapping sequence reads against a reference genome. Widely used alignment software tools include BWA (H. Li and Durbin 2009; H. Li 2013), Novoalign (www.novocraft.com) and Bowtie2 (Langmead et al. 2009). Although these utilities perform the same basic function, each relies on a unique set of algorithms and procedures, often resulting in somewhat different outcomes or performance. For example, BWA use the Burrows-Wheeler transform algorithm, while Novoalign uses the Needleman-Wunsch algorithm (H. Li and Durbin 2009; Thankaswamy-Kosalai et al. 2017).

Parts of this chapter were published as Wickland et al, 2017: *A comparison of genotyping-by-sequencing analysis methods on low-coverage crop datasets shows advantages of a new workflow, GB-eaSy*. BMC Bioinformatics 18(1): 586.

Studies comparing multiple aligners have shown that Novoalign achieves the greatest accuracy, but BWA is almost as accurate, considerably faster and less computationally costly (H. Li 2013; Thankaswamy-Kosalai et al. 2017).

A standard step after alignment is recalibration of base quality scores to correct systematic bias in base-quality computation by the sequencer. Perhaps the most widely used tool for recalibration is the Base Quality Score Recalibration (BQSR) algorithm implemented in the Genome Analysis Toolkit (GATK) software (Van der Auwera et al. 2013). BQSR creates a model of covariation for base quality scores (e.g. based on machine cycle, original quality score, dinucleotide content, location of nucleotide in the read, etc.) and adjusts quality scores to account for this variation.

Mapped reads with recomputed quality scores serve as input for variant calling, the identification of sampled sites polymorphic to the reference genome sequence. Most current variant calling tools, such as GATK HaplotypeCaller (Van der Auwera et al. 2013) and SAMtools (H. Li et al. 2009), use Bayesian statistical analysis of the base calls and quality scores to identify variant sites (Nielsen et al. 2011; O'Rawe et al. 2013). After variant calling, the process of genotype calling, which some tools combine with variant calling in one step, involves determining the consensus genotype in a group of reads from a particular sample and whether that consensus varies from the reference. Software programs for variant calling function on a per-sample basis, but genotype calling may be conducted either on a per-sample or joint basis. Joint genotyping achieves higher accuracy by leveraging information from all sequenced individuals to determine the genotypes at the variant positions identified (https://gatkforums.broadinstitute.org/gatk/discussion/4150/should-i-analyze-my-samples-alone-or-together); this method can rescue less-confident genotype calls that would otherwise be discarded. After variant calling and genotyping, filtering steps are implemented to minimize the number of false positive variants in the dataset. Downstream applications for the filtered variants include genome-wide association studies (GWAS) – which statistically analyze the variants to find genetic loci correlated with a trait of interest – population genetics studies, and measurement of RNA expression levels.

The choice of bioinformatics tools to process genomic data greatly influences the final set of detected variants. Different alignment and variant calling software packages rely on

different statistical models and computational algorithms, resulting in a substantial portion of non-overlapping variants between workflows implemented on the same raw data (O'Rawe et al. 2013; Pirooznia et al. 2014; Wickland et al. 2017), especially for low-frequency variants and large cohort sizes (Ren et al. 2018). Despite the lack of consensus among workflows, comparison of the variant call sets to highly confident, "gold standard" reference variant calls indicates that most of the variants uniquely identified by individual methods are valid. These results suggest that the optimal strategy to capture as many true variants as possible is to combine multiple analytic approaches (O'Rawe et al. 2013; Ren et al. 2018; Wickland et al. 2017). This integration strategy is especially critical in the search for the low-frequency (1-5%) and rare (frequency <1%) variants because each method on its own delivers an incomplete call set.

ALZHEIMER'S DISEASE GENOMICS

OVERVIEW OF ALZHEIMER'S DISEASE

Alzheimer's disease (AD) is a progressive neurological disorder characterized by dementia, pathological protein aggregation, synaptic degradation, and neural atrophy. In the U.S. alone, it affects over 5 million people and its medical costs amount to $250 billion annually, figures projected to increase to 14 million people and $1 trillion by 2050 (Hebert et al. 2013; Alzheimer's Association 2018). An estimated 3% of people aged 65-74, 17% of people aged 75-84, and 32% of people aged 85 and older have AD (Hebert et al. 2013). As the most common cause of dementia, AD impairs virtually all aspects of cognition, especially memory. Clinical symptoms typically surface after age 65, but the underlying neuropathological lesions can begin to emerge years or even decades earlier (Sperling et al. 2011). Although symptoms may first appear as occasional forgetfulness or mild short-term memory impairment, over time they eventually progress to the severe cognitive deficits characteristic of dementia, in which deteriorated brain function interferes with the most basic tasks of living (Sperling et al. 2011; Alzheimer's Association 2018).

Clinical diagnosis of possible AD, probable AD and dementia relies on cognitive and behavioral evaluation, often supplemented by assessment of chemical or neuroimaging biomarkers (Hyman et al. 2013; Frisoni et al. 2017). However, only histological inspection of brain tissue at autopsy can conclusively reveal the cell loss and key neural pathologies required for definitive AD diagnosis: extracellular amyloid plaques and intracellular neurofibrillary tangles. Amyloid plaques consist of aggregated amyloid beta, while neurofibrillary tangles consist of aggregated tau protein. Amyloid beta, tau and their precursor proteins play important roles in neural structure and function, but abnormal protein processing and age-related degeneration of protein homeostasis networks compromise the brain's ability to counteract perturbations in these pathways (Pearson and Peers 2006; Jack et al. 2010; Penke et al. 2017).

Amyloid beta forms from cleavage of Amyloid Precursor Protein (APP), a transmembrane protein that functions in synaptic formation and maintenance, signal transduction, and neural homeostasis and survival (Penke et al. 2017). APP cleavage occurs through two pathways; ninety percent of APP undergoes processing by the primary pathway, which yields products with neuroprotective and damage repair functions, while the remaining 10% of APP reaches the secondary pathway that generates amyloid beta. The normal functions of amyloid beta, present only at low concentrations, include roles in lipid transport, neuronal excitability and homeostasis, and synaptic plasticity (Pearson and Peers 2006; C. Liu et al. 2013). However, amyloid beta at high concentrations aggregates into plaques that contribute to AD pathogenesis. These high concentrations of amyloid beta disrupt synaptic signaling, degrade neural structure, and impair neural metabolism. Amyloid beta toxicity depends in part on tau, the other protein implicated in AD neuropathology (Penke et al. 2017).

Tau is a microtubule-associated protein involved in assembly and stabilization of microtubules in neurons (Guo et al. 2017). Phosphorylation of tau protein controls its activity. Normal, non-pathogenic phosphorylation of tau occurs transiently at low levels and functions in neuroprotection (Iqbal et al. 2010). In AD, however, levels of tau hyperphosphorylation exceed those of normal neurons by 3-4x. This irreversible hyperphosphorylation triggers apoptosis and promotes the accumulation of tau into neurofibrillary tangles that destabilize microtubules, interfere with cell signaling, and disrupt cognitive function (Iqbal et al. 2010; Guo et al. 2017). These tangles, which are not readily degraded, spread throughout the brain in a predictable

progression (Braak and Braak 1991; Hyman et al. 2013); reminiscent of prions, they appear to propagate intercellularly to induce formation of additional aberrant tau conformations (Guo et al. 2017).

The formation of neurofibrillary tangles and amyloid plaques typically precedes clinical manifestation of AD, and the extent of these accumulations generally correlates with dementia severity (Braak and Braak 1991; Sperling et al. 2011; Hyman et al. 2013). However, substantial patient-to-patent variability complicates efforts to understand the relationship between these neuropathological changes and cognitive impairment. The appearance of Alzheimer's lesions may not coincide with clinical symptoms of dementia, particularly in the early stages of the disease; some individuals even remain cognitively normal until death despite postmortem detection of widespread AD lesions (Sperling et al. 2011; Hyman et al. 2013). Interestingly, certain lifestyle factors, such as advanced educational attainment and regular intellectual activity, may help build "cognitive reserve"; this compensatory mechanism buffers against intellectual impairment by recruiting alternate cognitive resources or coping strategies (Sperling et al. 2011; C. Liu et al. 2013). These findings and others suggest that deposits of amyloid beta and tau are necessary but not sufficient to cause AD symptoms. Finally, the AD lesions themselves also exhibit diversity in assembly state, length, and post-translational modifications (Hyman et al. 2013). These variabilities have impeded efforts to characterize the genetic basis of AD.

ALZHEIMER'S DISEASE GENOMICS AND MISSING HERITABILITY

Alzheimer's disease has a strong genetic component, with heritability estimated at 60-80% (Gatz et al. 2006; Ertekin-Taner 2007). Heritability, or the proportion of a trait's phenotypic variance determined by additive genetic factors, is estimated based on phenotypic correlation data from closely related individuals (e.g. twin studies) (Yang et al. 2017). The genetics underlying AD heritability differ between the two broad classes of AD: early-onset AD (EOAD) and late-onset AD (LOAD). EOAD, which comprises 1% of AD cases, emerges before age 65, while late-onset AD (LOAD), which comprises all other cases, emerges after age 65 (C. Liu et al. 2013). Key factors contributing to EOAD heritability are highly penetrant, pathogenic mutations

in Presenilin1 (*PSEN1*), Presenilin2 (*PSEN2*) and the APP gene, all of which encode proteins that influence amyloid beta processing (Ertekin-Taner 2007; C. Liu et al. 2013; Cuyvers and Sleegers 2016).

The genetics of LOAD are more complex. The most well-established AD risk gene, *APOE*, encodes Apolipoprotein, which functions in cholesterol transport and neural response to injury (C. Liu et al. 2013). Two non-synonymous single nucleotide polymorphisms (SNPs) – Rs429358 (https://www.snpedia.com/index.php/Rs429358) and Rs7412 (https://www.snpedia.com/index.php/Rs7412) – in *APOE* create three haplotypes at this locus, referred to as the alleles *APOE* Ɛ2, *APOE* Ɛ3 and *APOE* Ɛ4. These alleles generate three isoforms of the APOE protein – APOE2 (Cys112, Cys158), APOE3 (Cys112, Arg158) and APOE4 (Arg112, Arg158) – each with a different contribution to AD susceptibility (Richard et al. 1994). Accounting for 17-27% of AD heritability, the Ɛ4 allele of *APOE* is the single greatest genetic risk factor for LOAD (Lambert 2013; Ridge et al. 2013; Cuyvers and Sleegers 2016). The Ɛ4/Ɛ4 genotype greatly increases an individual's risk of developing AD; 40% of AD patients, but just 15% of the general population, carry this genotype (Rubinsztein and Easton 1999). Compared to the most common genotype (Ɛ3/Ɛ3), Ɛ4 in one copy raises AD risk by ~3x and in two copies by ~15x among Caucasians, with a fairly similar pattern in other ethnic groups (Farrer et al. 1997). At the neural level, both human and rat neurons expressing the Ɛ4 allele show elevated levels of amyloid beta and induction of tau phosphorylation (Shi et al. 2017; Wang et al. 2018). In contrast, the Ɛ2 allele exerts a neuroprotective effect, and in AD patients it apparently delays disease onset and attenuates amyloid beta deposition relative to Ɛ4 (Serrano-Pozo et al. 2015).

GWAS have been undertaken to identify additional genetic factors associated with AD risk. Besides *APOE*, over 20 other AD-associated genes have been detected by GWAS (Cuyvers and Sleegers 2016). Functionally consequential variants lying within the coding sequence or regulatory regions of these genes have been reported, among them the AD risk alleles in Cluster (*CLU*), complement component receptor 1 (*CRI*), *BIN1*, *CD33* and Sortilin-related receptor 1 (*SORL1*). Despite these advances, these studies account for only a small fraction of the AD heritability deduced from phenotypic data (Ridge et al. 2013; Cuyvers and Sleegers 2016). The AD-associated loci thus far identified by GWAS, together with *APOE* Ɛ4, collectively explain 28-57% of the disease's heritability (Cuyvers and Sleegers 2016). The remaining heritability is

considered “missing.” The limitation of GWAS in accounting for trait heritability represents a major obstacle towards genetic characterization of this complex disease. The prevailing consensus is that LOAD results from a combination of common variants and rare, large-effect variants, the latter of which may comprise a substantial portion of the missing heritability (Cuyvers and Sleegers 2016; Lord et al. 2014; Marouli et al. 2017; Patel et al. 2019; Ridge et al. 2013).

Current research has focused on the contribution of rare mutations to AD risk. The case of the GWAS-detected gene *SORL1* illustrates the relevance of rare mutations to AD. The SORL1 protein modulates amyloid beta expression by preventing cleavage of APP into amyloid beta and by channeling newly synthesized amyloid beta to the lysosome for breakdown (Andersen et al. 2016). Recent reports have discovered very rare variants in this gene associated with substantially elevated AD risk. One study found rare, damaging *SORL1* variants (minor allele frequency [MAF] < 1%) in some individuals with EOAD that increase AD risk by 5x (Nicolas et al. 2016). Another report profiled additional highly penetrant, very rare *SORL1* variants that confer even greater AD risk. In a whole-exome cohort of 640 cases and 1268 controls, the majority of the very rare (MAF < .01%) *SORL1* variants predicted (by functional annotation tools) to be strongly damaging appeared only in single individuals (Holstege et al. 2017). Individuals carrying these strongly damaging “singletons” were over 10x more likely to develop AD, and the five singletons resulting in protein frameshift or truncation were observed only in AD cases. In addition, AD cases carrying these singletons had an earlier age of onset (58.9) than cases lacking these singletons (65.1). Conversely, more common variants identified in this study showed no association with AD risk, even when predicted to be strongly damaging based on protein changes.

To identify rare alleles, a current trend in AD sequencing studies is the use of large cohorts, which are required not only to capture the low-frequency variants themselves but also to achieve sufficient statistical power to establish the variants’ significant association with disease. In the context of GWAS, statistical power refers to the probability of recognizing a true association between disease and a genetic variant (i.e., correctly rejecting the null hypothesis of no association) (Sham and Purcell 2014). The Alzheimer’s Disease Sequencing Project (ADSP)

was initiated to identify novel AD-related genetic variation, and the large size of its case-control, whole-exome dataset of more than 10,000 samples was intended to raise statistical power compared to previous studies (Beecham et al. 2017). Recent reports have identified novel AD risk variants in this dataset and have offered additional support for previously established loci. One study found 11 variants – 8 of them residing in the *APOE* region of chromosome 19 – with statistically significant association with AD, as well as 14 suggestively significant variants (Bis et al. 2018). Of these significant SNPs, three falling outside of the *APOE* region reached significance in replication cohorts. However, no significant SNPs lying in genes with novel AD association reached significance in replication cohorts. Another study used a non-statistical approach to analyze this dataset (Patel et al. 2019). Rather than conducting conventional statistical tests, the researchers simply identified SNPs with MAF < 0.1%; discarded those with low predicted impact on disease, allele count below 3, or no effect on amino acid sequence; and counted the number of case or control individuals carrying these SNPs. This approach identified 32 SNPs in 24 previously reported AD genes, including two SNPs (in *TREM2* and *NOTCH3*) that were fully penetrant in the individuals carrying them.

ADSP and other large-scale sequencing projects have sought to address the missing heritability issue by capitalizing on large-cohort datasets to capture variants that are rare. Despite some success for rare variants analysis, much of the genetic contribution to AD remains unknown. Given the complexity of AD genetics and variability in AD phenotype, disease onset and disease progression, it is plausible that important types of genetic variation have not been explored sufficiently. Synonymous SNPs represent one such form of variation.

SYNONYMOUS SNPS AND THEIR POTENTIAL ROLE IN AD

The degeneracy of the genetic code allows multiple codons to specify a given amino acid. Synonymous SNPs are nucleotide substitutions that encode the same codon as the reference, “wild-type” sequence. Association studies typically disregard these variants because they do not alter a protein’s amino acid composition. However, the unequal frequencies of synonymous codons throughout the genome, known as codon usage bias, suggests their susceptibility to forces of selection and their functional significance (Chamary et al. 2006; Drummond and Wilke 2008). Indeed, synonymous SNPs can impact mRNA stability, translation speed and protein

structure and function (Chamary et al. 2006; Spencer et al. 2012; Sauna and Kimchi-Sarfaty 2011). The two classes of mutations typically studied, non-synonymous and frameshift mutations, represent the largest proportion of rare alleles, which are presumably deleterious and subject to purifying selection (Im et al. 2018). Synonymous SNPs comprise the second most abundant class of rare alleles, outnumbering the rare alleles of both introns and untranslated regions; this further suggests that these so-called "silent" mutations are under selection.

Because of their slightly different nucleotide compositions, synonymous codons can introduce subtle changes to mRNA secondary structure that influence the molecule's stability, which in turn affects its rate of degradation and ability to initiate translation (Gu et al. 2010; Sauna and Kimchi-Sarfaty 2011). For instance, more stable mRNA is less readily degraded and less readily translated. In addition, codon usage bias is linked to the composition of the tRNA pool in many organisms, including humans; more abundant codons, particularly in highly expressed genes, generally correspond to more concentrated isoacceptor tRNA species that recognize those codons, which maximizes the speed and accuracy of translation (Ikemura 1985; Lavner and Kotlar 2005; Chamary et al. 2006; Drummond and Wilke 2008; Yu et al. 2015). By affecting the rate of translational elongation, codon usage also modulates co-translational protein folding, a process that influences protein function (Spencer et al. 2012; Yu et al. 2015). For example, in many highly expressed proteins, structurally important amino acid residues are encoded by translationally optimal codons, a relationship that reduces the likelihood of protein misfolding arising from ribosomal stalling (Zhou et al. 2009).

Synonymous SNPs have been discovered in genes that underlie human diseases ranging from breast cancer to Crohn's disease (Hunt et al. 2014). A synonymous SNP in the Tristetraprolin (*TTP*) gene, which functions in suppression of breast cancer tumorigenesis, reduces expression of TTP protein by hampering translational efficiency (Griseri et al. 2011). Similarly, two synonymous SNPs in high-temperature requirement A1 (*HTRA1*), a gene strongly linked to neovascular age-related macular degeneration, are over-represented in individuals afflicted with this condition. These synonymous SNPs, each involving a switch to a less frequently used codon, result in similar mRNA expression but reduced protein production and reduced catalytic activity relative to wild-type codons, apparently by slowing translation speed

and interfering with protein folding (Jacobo et al. 2013). Finally, a synonymous SNP in the cystic fibrosis transmembrane conductance regulator (*CFTR*), which coordinates the movement of ions across epithelial cell membranes, has been documented in individuals with cystic fibrosis (Kirchner et al. 2017). While still specifying the amino acid threonine, this T-to-G SNP has a low-abundance cognate tRNA in bronchial epithelial tissue. The low concentration of this tRNA retards translation, which alters co-translational folding and ultimately impairs ion channel conductance. These results establish that synonymous mutations can impact protein structure and function in disease-related genes.

In the scientific literature, coverage of synonymous SNPs with respect to Alzheimer's disease is sparse. One recent study found two genes containing synonymous SNPs correlated with entorhinal cortical thickness (an AD imaging biomarker) in individuals with AD (J. E. Miller et al. 2018). However, it is possible that these variants were simply markers in linkage disequilibrium with causative nucleotide changes or that the cortical thickness phenotype was not directly related to AD. Therefore, much remains unknown concerning the potential connection between synonymous variants and AD (see Chapter 2).

GENOMIC SEQUENCING ANALYSIS IN PLANTS

CHALLENGES IN PLANT GENOMICS

Certain biological characteristics of plant genetics complicate the bioinformatic analysis of plant genomes relative to humans and other animals. Challenges associated with plant genomes include polyploidy and high repeat content (Jiao and Schneeberger 2017). Polyploids are organisms that possess more than two sets of homologous chromosomes. Widespread in plants compared to animals and other organisms (Orr 1990), polyploidy can originate from whole-genome duplication in a single species (autopolyploidy) or, more commonly, from chromosome doubling after interspecific hybridization (allopolyploidy) (Kyriakidou et al. 2018). An estimated 80% of flowering plants are polyploid or have a polyploid lineage. Many extant plants, such as soybean (*Glycine max*), are paleopolyploids whose duplication events occurred

millions of years ago and that have since undergone partial diploidization, the return to diploid-like cytogenetic behavior as homeologs diverge (Blanc 2004; Schmutz et al. 2010; Walling et al. 2006; Wolfe 2001; Kim et al. 2009). Polyploidy acts as a source of variation upon which evolution may act to introduce novel adaptations. For example, duplication of an ancestral flowering gene, likely from a whole-genome duplication event during angiosperm evolution (Y. Liu et al. 2016), led to diversification of function into separate clades with opposing roles in flowering initiation (reviewed in Wickland & Hanzawa, 2015). Polyploidy also buffers against the impact of deleterious mutations. Although useful biologically, the presence of closely related homeologous chromosomes presents challenges to mapping algorithms; during alignment, reads may be mapped to the wrong homeolog. In addition, failure to distinguish polymorphisms between sub-genomes ("homeoSNPs") from true allelic SNPs further raises the error rate in the set of detected variants (Clevenger and Ozias-Akins 2015).

A related characteristic of plant genomes that poses additional challenges is high repeat content, which is particularly common in plants (Nicholas et al. 2016). Repetitive sequences, especially long repeat sequences not fully spanned by short reads, impede correct read placement against the reference, raising the error rate (Claros et al. 2012; Treangen and Salzberg 2012). Leading contributors to high repeat content in plants are transposable elements, which are rich in repetitive sequences and comprise over half of the genomes of soybean and several grasses and over 85% of the maize genome (Claros et al. 2012; Nicholas et al. 2016; Feschotte et al. 2002; Gao et al. 2012; Schmutz et al. 2010). High repetitiveness also influences assembly of the reference genome itself, and specialized assembly tools have been developed to handle the repetitiveness inherent in plant genomes (Bolger et al. 2017). In addition, several methods have been advanced that can help mitigate issues related to high repeat content in reference-based variant calling. For example, paired-end reads, which represent sequence from both ends of a DNA fragment, can facilitate correct read placement relative to the reference genome; each read in a pair may contain flanking sequence adjacent to a different side of the repetitive region. A related, and more recent, method is long-read sequencing technology, which produces reads over 10 kilobases in length (Jiao and Schneeberger 2017; Rhoads and Au 2015). Advanced by companies such as Pacific Biosciences and Oxford Nanopore, this technology generates reads long enough to "sequence through" extensively repetitive regions;

however, per-nucleotide error rates remain relatively high compared to short-read sequencing and costs are usually much higher, making short-read sequencing the technology of choice for most variant detection applications.

In addition to biological factors, financial considerations impact the analysis of plant genomes relative to those of humans. In 2019, the budget of the National Institutes of Health (NIH), the medical-research arm of the United States Department of Health and Human Services, exceeded $39 billion, more than 80% of which was distributed as competitive grants to researchers (https://www.nih.gov/about-nih/what-we-do/budget). In contrast, the United States Department of Agriculture (USDA) research budget in the same year amounted to $2.6 billion; approximately half of this funded the research activities of the Agricultural Research Service (ARS) and the National Institute of Food and Agriculture (NIFA), the latter awarding $375 million in competitive grants (United States Department of Agriculture, 2019). Although other sources of research funding exist, the large disparity in research budgets between the NIH and USDA reflects the reduced funding available for plant research compared to human/medical research at large. For this reason, many of the methods developed specifically for plant research aim to reduce costs wherever possible.

GENOTYPING BY SEQUENCING (GBS)

Whole-genome and whole-exome sequencing can identify millions of SNPs, but for many applications involving genetic linkage in plants, such high densities of markers are unnecessary and costly. Reduced-representation approaches involve sequencing a subset of locations spread throughout the genome to reduce genome complexity, minimize costs and rapidly genotype samples using SNP markers. The earliest developed reduced-representation sequencing method, restriction site associated DNA (RAD) sequencing, uses restriction enzymes to divide the genome into sheared DNA fragments, which are size fractionated and then sequenced on NGS platforms (Baird et al. 2008; M. R. Miller et al. 2007; Scheben et al. 2017). Often, restriction enzymes are selected that cut infrequently in repetitive regions in order to reduce the likelihood of obtaining reads that map to multiple locations. RAD sequencing remains the method of choice for biological diversity applications in which reference genomes are not available. In this and similar methods, each sample is assigned a unique barcoded adapter for multiplexed

sequencing in a single Illumina flow-cell lane, thereby increasing the number of samples under investigation and reducing financial costs. Although this method works well on crops such as soybean (Varala et al. 2011), the large amount of high-quality DNA required for the size selection step, and consequent higher DNA preparation costs, makes RAD sequencing unsuitable for routine use in plant breeding.

Genotyping-by-sequencing (GBS), a simplified reduced-representation sequencing approach (Elshire et al. 2011), has gained popularity in crop research and plant breeding for high throughput, low-cost genotyping. It has been applied to projects ranging from genomic selection to gene mapping to GWAS in numerous crop species (Furuta et al. 2017; H. Liu et al. 2014; Poland et al. 2012; Sonah et al. 2015; Wu et al. 2016). Like RAD sequencing, GBS relies on restriction enzymes to generate a reduced representation of the genome for sequencing. However, the GBS library preparation protocol involves fewer steps than RAD sequencing, requires less DNA, and lacks a size selection step (Elshire et al. 2011). In GBS, DNA samples are digested and ligated to barcoded adapters in single wells, pooled, and then enriched by PCR.

Bioinformatics software packages and workflows have been developed to facilitate analysis of reduced-representation sequencing data (Catchen et al. 2013; Elshire et al. 2011; Sonah et al. 2013; Torkamaneh et al. 2017). Several of these platforms utilize the same tools and algorithms commonly applied to whole-genome sequence data, while others utilize algorithms developed specifically for GBS and RAD sequencing. Comparisons of GBS analysis methods show that a substantial portion of detected variants are unique to each bioinformatics tool (Sonah et al. 2013; Torkamaneh et al. 2016; Wickland et al. 2017). Contributions to this low overlap include the different statistical approaches used by each method to determine the consensus genotype in a group of reads and whether that consensus differs from the reference sequence. These methodological differences, combined with the polyploid nature of plant genomes and the large proportion of missing data inherent in GBS due to the reduced-representation approach, likely account for the lack of consensus between tools. Despite relatively low convergence between tools, SNP calls generally show high accuracy based on validation using other methods, indicating that different GBS analysis methods arrive at largely complementary sets of valid SNP calls. Therefore, a comprehensive approach integrating the results of multiple

bioinformatics pipelines may be a key strategy to obtain the largest, most highly accurate SNP yield possible for reduced-representation, low-coverage sequencing data (Wickland et al. 2017; see Chapter 3).

CHAPTER 2: THE IMPACT OF BATCH EFFECT AND STUDY DESIGN ON IDENTIFICATION OF GENETIC RISK FACTORS IN HUMAN SEQUENCING DATA

BACKGROUND

Genetic studies have shifted from SNP array-based genome-wide association study to rare variants discovery by exome and whole-genome sequencing. Sequencing-based searches for rare disease-associated variants require large sample sizes to achieve sufficient statistical power, but they often entail batch effects and biases from study design. Batch effects refer to sources of variation arising not from the targeted biological differences between phenotype classes but from differences between experimental or technological batches. If not adequately addressed in the analysis, batch effects reduce statistical power and raise susceptibility to false-positive associations, and biases in study design may further hinder the ability to detect true genotype-trait associations. A standard practice in association studies is to use statistical models adjusted for batch effects and other heterogeneity in the dataset, followed by additional quality control of the identified genetic risk variants.

Practices that may introduce batch effects include dividing samples among multiple sequencing centers, collecting samples under different protocols, and extracting exomes using different target capture kits. For example, the Alzheimer's Disease Sequencing Project (ADSP) sequenced exomes of more than 10,000 cases and controls to identify genetic variations associated with Alzheimer's disease (AD) (Beecham et al. 2017). Sequencing for this dataset took place at three centers: the Broad Institute (Broad), the McDonnell Genome Institute at Washington University (WashU), and the Human Genome Sequencing Center at Baylor College of Medicine (Baylor). Broad prepared sequencing libraries using the Illumina Rapid Capture Exome kit, while WashU and Baylor used the Roche Nimblegen VCRome v2.1 kit (https://www.niagads.org/adsp/content/sequencing-pipelines). In addition to three sequencing centers and two different exome capture kits, another potential confounding factor is the age distribution of sample classes, with cases averaging approximately 11 years younger than controls. The intentional selection of older controls was designed to identify AD-causal variants

that are absent from older but cognitively normal individuals, but this lack of independence between age and case-control status could also confound the association analyses.

We studied batch effects and confounding variables in the ADSP dataset and found that both impacted the association analyses. In particular, we identified significant differences in genotype quality and allelic capture biases between the two exome capture kits. In addition, we investigated the influence of age and allele frequency differences between sequencing center cohorts on variant association with AD.

METHODS

Dataset description

The Sequence Read Archive (SRA) files containing the raw sequencing data of 10,993 AD cases and controls were downloaded from dbGap and converted to FASTQ format using the SRA Toolkit (https://www.ncbi.nlm.nih.gov/books/NBK158899). Access to this public dataset was approved by the Institutional Review Board (IRB) of Mayo Clinic, the IRB of the University of Illinois, and dbGAP (https://www.ncbi.nlm.nih.gov/gap/). The Alzheimer's cases satisfied the National Institute on Aging and the Alzheimer's Association criteria (McKhann et al. 2011) for definite, possible or probable Alzheimer's disease. These cases included patients with and without *APOE* (Corder et al. 1993) risk alleles. The controls were at least 60 years old, showed no sign of dementia based on cognitive testing, and scored low on risk assessment (Beecham et al. 2017). Of the 10,993 samples, 9,904 passed sample-level quality control based on the following criteria: (1) variant call rate $\geq$ 95% per sample; (2) coverage $\geq$ 10x for at least 90% of exome; (3) *APOE* genotype match between cohort meta-data and sequenced genotypic data; (4) average transition/transversion ratio $\geq$ 2.8; (5) FREEMIX (Jun et al. 2012) sample contamination estimate > 0.02 ; (6) gender error PLINK F estimate < 0.07 for males and > 0.03 for females. Ancestry of 99.8% of samples was classified as European.

Variant calling

The paired-end sequence reads were aligned to the human reference genome build 37 using Novoalign (http://www.novocraft.com) (default parameters), which was selected on the

basis of its greater accuracy in read placement relative to other methods (H. Li 2013; Thankaswamy-Kosalai et al. 2017) and its lack of prior application to this dataset for association testing (e.g. Bis et al. 2018; Patel et al. 2019). The alignment files were then sorted by read position using Novosort (http://www.novocraft.com), realigned around small insertions and deletions (INDELs) using Picard (https://broadinstitute.github.io/picard/), and subjected to base recalibration using the Genome Analysis Toolkit (GATK) version 3.4 (Van der Auwera et al. 2013). Variant calling followed GATK's best practices guidelines for germline variants (https://software.broadinstitute.org/gatk/best-practices/workflow?id=11145): per-sample variant calling on the realigned, recalibrated BAM files was performed using HaplotypeCaller, and multi-sample joint genotyping of all 9,904 samples was performed using GenotypeGVCFs. Variant calling was conducted only on the exome regions common between the two exome capture kits (Illumina Rapid Capture Exome kit and Nimblegen VCRome v2.1 kit). Variants were annotated by snpEff (Cingolani et al. 2012) and ANNOVAR (Wang et al. 2010). All data processing was carried out on the Blue Waters supercomputer at the University of Illinois at Urbana-Champaign.

Variant-level quality control

Several steps were undertaken to minimize the number of false-positive variant calls. The Variant Quality Score Recalibration (VQSR) step implemented in GATK uses machine learning algorithms to compute new, well-calibrated quality scores for all variants based on the annotations of a high-quality subset of the analyzed data. In accordance with GATK Best Practices for whole-exome data, the variables included in the VQSR model consisted of QD, MQ, MQRankSum, ReadPosRankSum, FS, SOR and InbreedingCoeff for SNPs; and QD, MQRankSum, ReadPOsRankSum, FS, SOR and InbreedingCoeff for INDELs (Van der Auwera et al. 2013). A sensitivity threshold of 99.5 was used for SNPs and 99.0 for INDELs. Detected variants were excluded from further analysis if they failed VQSR, deviated significantly ($p < 1.0 \times 10^{-6}$) from Hardy-Weinberg equilibrium (HWE) in the control samples, or had an alternate allele call supported by less than 10 reads across the cohort.

Association tests and statistical models

After quality control, association testing using disease status (case or control) as the phenotype was performed on the variants under four additive logistic regression models implemented in Plink 1.9 (Chang et al. 2015). Each model included a unique combination of the following covariates: sequencing center, sex, age, *APOE* genotypes, and the first four principal components (PCs) underlying population structure (**Table 2.1**). All models adjusted for sequencing center and the first four PCs. Model 1 used only sequencing center and the first four PCs as covariates, leaving out *APOE* genotype and age because AD cases consisted of patients with and without *APOE* risk alleles and because age confounds with AD status. Model 2a included *APOE* genotypes as a covariate; Model 2b adjusted for age but left out *APOE*; and Model 3 adjusted for all listed covariates. The association tests were conducted on all 9,904 samples together as well as on sets of samples stratified by sequencing center and age. Variants were considered exome-wide statistically significant at the Bonferroni-corrected threshold of $p < 0.05 / \#\text{ tests}$ and suggestively significant at $p < 1 / \#\text{ tests}$ (e.g. Bis et al. 2018).

Principal components analysis

Population substructure

PCs of the detected genotypes were calculated after excluding variants that failed VQSR; variants with a call rate below 95%, minor allele frequency (MAF) below 5%, or HWE deviation below $p < 1.0 \times 10^{-5}$; variants lying within the highly variable HLA, LCT, 8p and 17q regions; and variants with a linkage disequilibrium r^2 value above 0.2. Only the first four PCs had eigenvalues above 1, so these were retained in the statistical models for association testing to adjust for population substructure.

Visualization of batch effect

To explore the possibility that a batch effect originated from the use of two capture kits, we compared genotype-level quality metrics between the Broad cohort (Illumina kit) and the WashU and Baylor cohorts (Nimblegen kit). Quality metrics under consideration were genotype quality (GQ), read depth (DP) and minor allele

concentration (MAC). GQ measures the variant caller's confidence in assigning a genotype at a given SNP for a given sample. DP refers to the number of sequenced reads that support a given genomic position for a given sample. MAC refers to the proportion of sequenced reads that support the minor allele, and is calculated by dividing the number of reads supporting the minor allele by the total number of reads at a given position for a given sample.

Mean values for these three metrics at each variant position were computed separately for the Illumina-captured (Broad) samples and the Nimblegen-captured (WashU and Baylor) samples. The ratio of the means for each kit was taken at each SNP for each quality metric. Variants with large discrepancies between the two capture kits reside at the tails of the distribution of these ratios. To ascertain whether these discrepant variants could differentiate the genotypes detected by each capture kit, PCs were computed using as input the 9,904 samples' genotypes at the SNPs lying within 1) the outer 10% of each quality-metric ratio distribution and 2) the middle 90% of each distribution.

RESULTS

Description of detected variants

Of the 1,584,609 variants detected across the cohort, 166,947 variants passed VQSR and the additional filtering steps detailed above. These variants totaled 120,572 from the Broad samples; 108,390 from the WashU samples; and 98,542 from the Baylor samples. Approximately 70% of variants were shared among samples from all three sequencing centers (**Figure 2.1**). The larger number of variants detected in Broad samples is likely due to the larger number of individuals sequenced by Broad compared to the other two centers (**Table 2.2**).

Association analysis

Full-cohort association analysis

We first applied the four models (**Table 2.1**) to the full dataset of 9,904 samples. As shown in Figure 2.2, Models 2b and 3 identified very few significant variants, as expected, because age confounds with AD status. Under Model 1, which adjusted only for sequencing center and PCs, 73 variants reached exome-wide significance ($p < 3.0 \times 10^{-7}$) and an additional 27 variants reached suggestive significance ($p < 6.0 \times 10^{\wedge -6}$) (**Figure 2.2**). The most highly significant SNPs under this model reside on chromosome 19, which contains well-established AD risk alleles and protective alleles in the gene *APOE*. Model 2a, which accounted for sequencing center and PCs as well as sex and *APOE*, found 43 significant variants and 48 additional suggestive variants (**Figure 2.2**). Most of these 43 SNPs and their corresponding genes had no previously reported association with AD (e.g. Bis et al. 2018). In addition, all significant variants detected under Model 2a were also significant under Model 1 (**Figure 2.3**). To characterize batch effects and other confounding variables, we focused on the 43 SNPs attaining exome-wide significance under both Model 1 and 2a. These 43 SNPs were further filtered to remove multi-allelic SNPs, INDELs, and any SNP with a recalibrated variant quality score (VQSLOD) < 0, resulting in a set of 30 top SNPs for additional analysis (**Table 2.3**).

The significance of AD association came from Broad samples only

To investigate batch effects and other variables related to sequencing center, we repeated the four association models separately on the Broad, WashU and Baylor cohorts. Unexpectedly, at 28 of the 30 SNPs identified as exome-wide significant in the full-dataset analysis, only the Broad cohort remained significant (**Figure 2.4**). Two of the 30 SNPs failed to reach significance in any individual cohort. Consistent with the observation of Broad-exclusive significance, the minor allele frequency (MAF) of the 28 SNPs showed clear differences between cases and controls only in the Broad samples (**Figure 2.5**). These findings indicate that Broad cases drove the significance observed in the full cohort at these SNPs. This sequencing center difference was observed not only

at the top 30 SNPs; we examined all nominally significant AD variants ($p < 0.005$) and saw vast differences between sequencing centers (**Figure 2.6**). Under Model 1, only 8 out of 1,730 nominally AD-associated ($p < 0.005$) variants were shared across all three centers; Models 2a, 2b, and 3 identified very few or zero AD variants shared across centers.

Factors underlying the discrepancies between sequencing centers

Genotype-level quality metrics by capture kit and sequencing center

The surprisingly low overlap and incongruent significance of AD-associated variants across sequencing center cohorts prompted us to further examine the characteristics of the detected variants across these centers. Since Broad used a different exome capture kit than the one used by WashU and Baylor, we compared the genotype quality (GQ), read depth (DP) and minor allele concentration (MAC) at all detected SNPs (**Figure 2.7**). As shown in Figure 2.7a-b, the log-adjusted distributions of the GQ and DP ratios between the exomes captured by the Illumina kit (Broad) vs. Nimblegen kit (WashU and Bylor) are normal, with the tails corresponding to SNPs with bigger differences in GQ and DP between two capture kits. Interestingly, the top 10% (5% from each tail) of SNPs most discrepant in GQ, but not DP, between two exome kits contributed to the batch effects between sequencing centers; PCs computed based on these SNPs show clear separation by sequencing center for GQ (**Figure 2.8**). For MAC, Broad samples had substantially lower values consistent with biases of less efficient capture for the alternate allele by the Illumina exome capture kit, as indicated by the left-skewed distribution in Figure 2.7c and the PCs in Figure 2.8. In contrast, the middle 90% of all three distributions showed no visual batch effect between sequencing centers (**Figure 2.9**). These results may indicate that global differences in GQ and MAC related to capture kit contributed to batch effects in this dataset.

To assess the possibility that the disparate allele frequencies among sequencing center cohorts arose from sample-level quality issues, we next profiled for each center the distributions of GQ, DP and MAC of the 30 SNPs significant from the full-dataset

analysis. Possible GQ values range from 0 to 100, with 30 and above generally considered highly confident. Based on this metric, the 30 SNPs exhibited high quality in both cases and controls in the Broad, WashU and Baylor cohorts (with the possible exception of 19:42799299 in the gene CIC) (**Figure 2.10**). No consistent trends in GQ emerged between cases and controls or between samples from different centers. Although the distributions of DP varied widely among SNPs, for a given SNP they were generally similar across sequencing centers (**Figure 2.11**). MAC is expected to be near 50% because (in theory) approximately 50% of reads should support the alternate allele in a heterozygous individual. However, in practice MAC often has a lower value due to different sequencing affinities of the alleles. Similar to those of GQ and DP, the distributions of MAC showed no clear pattern of differences among sequencing centers (**Figure 2.12**). Therefore, sample quality at these 30 SNPs appeared relatively uniform across the full cohort. In addition, all but 4 of the 30 SNPs fell in the middle 90% of each global annotation ratio distribution described earlier, suggesting that quality issues did not fully explain the sequencing center differences manifested in the discrepant minor allele frequencies of the 30 SNPs profiled and that other unknown factors also contributed to batch effect in this dataset.

Age distributions of cases and controls in Broad samples

Next, we examined the reasons why exclusively the Broad-sequenced samples yielded 30 highly significant AD SNPs under Models 1 and 2a. As shown in Table 2.2 and Figure 2.13, Broad sequenced the largest ADSP cohort, including 48.7% of all cases and 39.6% of all controls, indicating greater statistical power. In addition, Broad cases and controls had the biggest difference in average age, 13 years, compared to 8.5 and 10.5 years for WashU and Baylor, respectively. Furthermore, the distribution of Broad cases is shifted towards younger individuals compared to the other two centers; we divided the samples into four age groups (<64, 65-74, 75-84 and 85+) and observed that Broad-sequenced cases were over-represented in the younger age groups compared to WashU and Baylor (**Figure 2.14**).

Association analysis stratified by sequencing center and age

To examine the contribution of Broad's over-representation of younger cases to center-specific AD associations, we conducted the association tests separately on each of the four age groups listed above. Under this stratification scheme, the AD association fell to levels below suggestive significance in all age groups except age 85+ (**Figure 2.15**). In Broad samples aged 85 and older, 4 of the 30 SNPs attained exome-wide significance and an additional 4 SNPs attained suggestive significance. However, samples sequenced by WashU and Baylor failed to approach significance at any SNP. Intriguingly, the exome-wide significance in the oldest group at some of these SNPs in Broad may be due to a reverse of the MAF in cases vs. controls between individuals below or above age 74 (**Figure 2.16**). In Broad samples 74 years of age and younger, the MAF of nearly all of the top 30 SNPS was higher in controls compared to cases. However, in Broad samples older than 74 years, the MAF of nearly all top 30 SNPs was higher in cases compared to controls, which may have contributed to the significant p-values of these risk variants. In addition, both cases and controls sequenced by Broad showed declining MAF with age, suggesting that the SNPs are associated with age in these samples. These observations may indicate that the AD association detected in the full-dataset analysis was influenced by high MAF in both the old and young Broad cases. In contrast to the 30 statistically significant SNPs under both Model 1 and Model 2a, SNPs in 2 known AD genes (*APOE* and *TOMM40*) significant only under Model 1 showed similar MAF across all three centers and clear separation between case and control MAF (**Figure 2.17**).

DISCUSSION

To characterize the impact of batch effect and study design on downstream genomic analysis, we conducted association tests on the 9,904 exomes of the ADSP case-control dataset, which is composed of cohorts from three sequencing centers. We used four models with different sets of covariates. Each model adjusted for PCs and sequencing center. Model 1, the base model, and Model 2a, which further adjusted for *APOE*, identified a shared set of 30 highly

significant SNPs after quality control. Most of these SNPs had no previously reported AD association based on a review of the literature. In contrast to Models 1 and 2a, the two models that accounted for age (Models 2b and 3) found few SNPs statistically significant, indicating that the adjustment applied for age (a known confounding variable, as described above) eliminated the association between sample class (case or control) and disease. This result echoes a recent report that found only one exome-wide significant SNP in the ADSP dataset under age-adjusted models (Bis et al. 2018).

The impact of batch on the detected associations

The fact that multiple sequencing centers and exome capture kits were used to create this dataset prompted us to investigate batch effects and their impact on association analysis. We repeated the association tests separately on each center's set of samples and found that just 8 SNPs significant under Model 1 and 1 SNP significant under Model 2b were shared across all three cohorts; zero were shared between all three cohorts under Models 2a and 3. We also investigated cohort-level differences in the 30 highly significant AD-associated SNPs identified in the full-dataset analysis. Although not significant in the WashU and Baylor cohorts, these 30 SNPs displayed highly significant AD association and relatively high MAF in the Broad cohort, mirroring the results seen in the full-dataset analysis. This result indicates that the Broad samples drove the significance of these SNPs in the full-dataset analysis despite the use of a covariate to adjust for sequencing center, perhaps due to the complex relationship between center, age and AD association.

Several distinctions between the cohorts assembled at each center could have influenced the paucity of shared AD-associated variants. One such distinction is the exome capture kit used by each center. The comparison of the quality-metric ratio distributions between the Illumina-captured cohort (Broad) and the Nimblegen-captured cohorts (WashU and Baylor) revealed that variants with the greatest disparities in GQ could distinguish the genotypes detected by the two capture kits. Therefore, systematic differences in GQ between capture kits could contribute to the minimal overlap of significant variants across cohorts. However, 26 of the 30 of the highly significant SNPs from the full-dataset analysis without age correction did not exhibit large differences in GQ or other metrics between the Broad, WashU

and Baylor samples. Moreover, GQ was generally high at these SNPs. These results indicate that quality issues alone could not explain the differences between centers manifested in the discrepant minor allele frequencies of the 30 SNPs profiled, although GQ may well have contributed to the differences between capture kits at a more global level. Instead, a more likely explanation for the observed trends of significance and MAF at the 30 SNPs involves heterogeneity among the sampled individuals assembled for each cohort.

The impact of age and population stratification on the detected associations

We speculated that age differences may have contributed to the observed discrepancies in significance between sequencing centers because the Broad cohort consisted of a disproportionately large number of younger cases. To understand how this disparity might influence the ability to detect AD-related variation, we repeated the association tests on four age groups within each center's cohort and examined the significance of the 30 SNPs profiled earlier. Under this analysis, their significance fell to negligible levels in all age groups except the oldest (85+) in Broad. At 4 of the 30 SNPs, the Broad samples reached exome-wide significance in samples aged 85 and older, which implies true AD association of these SNPs in this age group in Broad. However, the remaining 26 SNPs were not statistically significant in any age group in Broad under this stratified analysis, possibly due to reduced power from smaller sample size within each age stratum. The 30 SNPs were more frequent among all age groups in Broad cases and controls compared to WashU and Baylor, particularly in younger samples. The parallel decline in MAF with age in both Broad cases and Broad controls at all 30 SNPs suggests that the SNPs are associated with age in Broad. Although the over-representation of younger Broad individuals classified as cases may have impacted the detected AD association, the impact of age does not exclude the possibility of true AD association as well.

The biological reason for higher MAF in Broad samples compared to WashU and Baylor samples at these SNPs remains unclear from the available data. One possibility is divergent genetic backgrounds among the sub-populations sampled in each center's cohort, although more information is needed to investigate this scenario. Population stratification, which is the regular pattern of disparate allele frequencies between sub-populations, reflects ancestral

differences in study participants (M. Li et al. 2010). To account for differing genetic backgrounds, association models generally include a covariate that adjusts for PCs of the detected genotypes, which are often correlated with ancestry (Martin et al. 2018). However, the global ancestral differences captured by PCs may not reflect sub-population differences at the individual SNP level, which may explain why the PC covariate failed to reduce the significance of the 30 profiled SNPs to lower levels.

The possible biological bases for the significance of the top SNPs

Regardless of the origin of their Broad-specific association with AD and age, several of these SNPs lie in genes previously linked to neural diseases: *TMED1*, *PLXNB1*, *CIC*, *CEP164* and *CCNK*. TMED1 interacts with APP (Del Prete et al. 2016), whose cleavage into amyloid beta generates one of the key components of AD-associated pathological protein aggregation (Penke et al. 2017). PLXNB1 influences amyloid beta load (Mostafavi et al. 2018), and CIC is a transcriptional repressor whose inactivation promotes gliomagenesis, the formation of glial tumors in the brain (Yang et al. 2017). CEP164 binds to TTBK2, a kinase that phosphorylates tau (Liao et al. 2015) and that contributes to neurodegeneration in frontotemporal dementia (Taylor et al. 2018).

Interestingly, we found two statistically significant synonymous SNPs and one suggestively significant synonymous SNP in *CCNK*, but no significant non-synonymous SNPs in this gene. Mutations in *CCNK*, which encodes the transcriptional regulator Cyclin K, have been previously associated with neurodevelopmental abnormalities (Fan et al. 2018). Synonymous SNPs are typically excluded from GWAS analysis because they do not change amino acid composition; however, differences between synonymous codons can impact mRNA secondary structure (Gu et al. 2010), the rate of translation (Ikemura 1985; Spencer et al. 2012) and co-translational protein folding (Spencer et al. 2012) and have been reported in connection to disease (reviewed in Hunt et al. 2014). Given the role of this class of variants in protein dynamics and its customary exclusion from GWAS analysis, the potential contribution of synonymous SNPs to disorders characterized by pathological protein aggregation, such as AD, has been overlooked and warrants further study.

Although not conclusive, these intriguing connections between neural disease and the SNPs identified here indicate that there may be a biological basis of the observed findings in the Broad cohort, but it is not possible from these data to determine whether the SNPs are associated with AD or with longevity due to other factors. Further study is necessary of both the population genetics of these SNPs in Broad and their impact on aging and AD.

CONCLUSIONS

To illustrate the impact of batch effect and confounding variables on downstream analysis, we conducted stratified association analysis on AD exome data using models with multiple combinations of covariates. We profiled a set of SNPs with highly significant, novel associations with AD that were impacted by heterogeneity in sub-cohort composition. We identified genotype quality, age and population stratification as likely contributing factors to vastly different minor allele frequencies across sequencing center cohorts. Collectively, our findings suggest that exome sequencing and other studies should follow consistent sample collection and sequencing protocols, use the same target capture kit, and minimize variation unrelated to disease phenotype between sample classes in order to locate truly significant disease-associated loci for challenging diseases.

ACKNOWLEDGMENTS

This research was supported in part by the Mayo Clinic and Illinois Alliance Fellowships for Technology-Based Healthcare Research program. This research is also part of the Blue Waters sustained-petascale computing project, which is supported by the National Science Foundation (awards OCI-0725070 and ACI-1238993) and the state of Illinois. Blue Waters is a joint effort of the University of Illinois at Urbana-Champaign and its National Center for Supercomputing Applications.

FIGURES AND TABLES

	COVARIATES				
	Seq. facility	PCs	Sex	Age	ApoE
Model 1	■	■			
Model 2a	■	■	■		■
Model 2b	■	■	■	■	
Model 3	■	■	■	■	■

Table 2.1. Covariates included in each model for association tests.

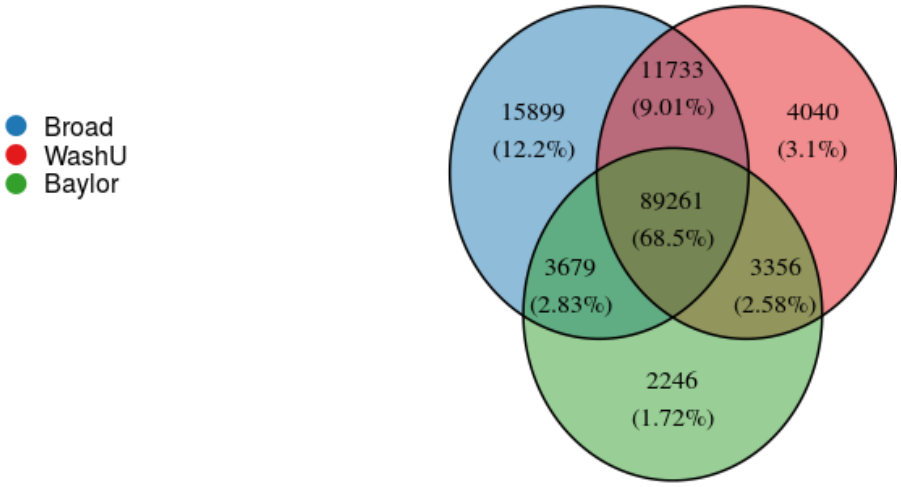

Figure 2.1. Number of QCed variants and their overlap among samples from three sequencing centers.

	All samples		Broad samples		WashU samples		Baylor samples	
	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls
No. of samples	5518	4386	2687	1740	1603	1657	1228	989
Freq. in cohort	55.7%	44.3%	27.1%	17.6%	16.2%	16.7%	12.4%	9.99%
Mean age	75.35	86.59	73.69	86.94	77.94	86.51	75.62	86.13

Table 2.2. Sample counts and mean ages of cases and controls for each sequencing center

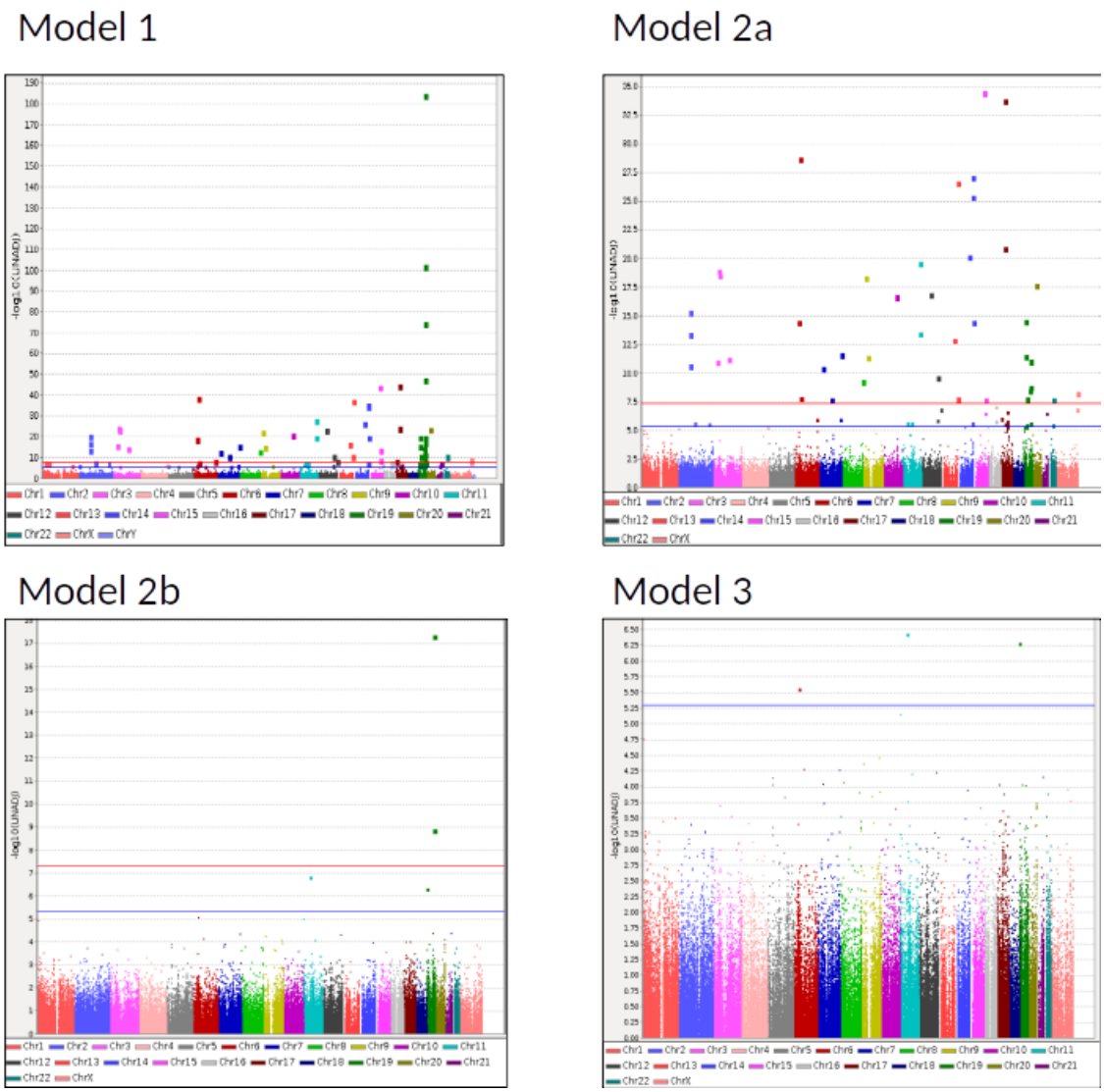

Figure 2.2. Manhattan plots of AD association under four models. Log-transformed p-values under each model are shown. All models included the covariates of sequencing center and PCs. In addition, Model 2a included *APOE* and sex, Model 2b included age and sex, and Model 3 included *APOE*, age and sex. The red horizontal line denotes exome-wide significance and the blue horizontal line denotes suggestive significance.

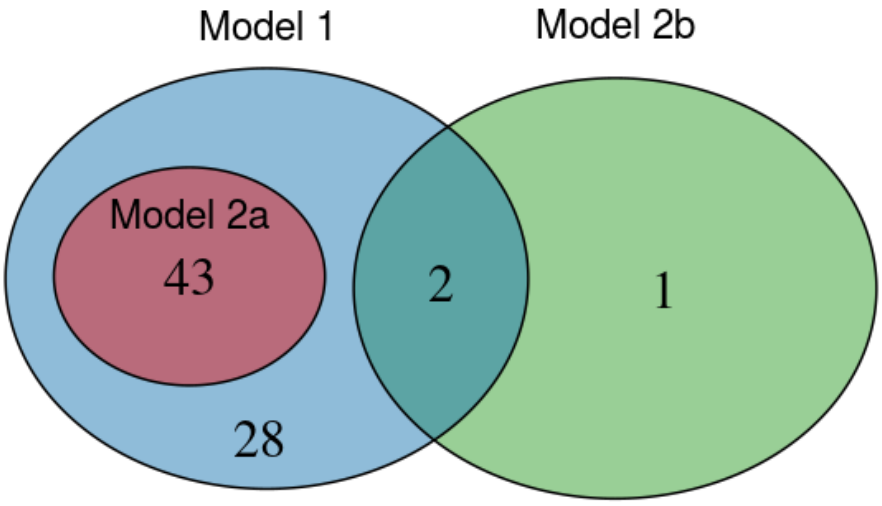

Figure 2.3. Number of statistically significant ($p < 3.0 \times 10\text{-}7$) AD-associated variants under three models. No significant SNPs were detected under Model 3.

Chr	Position	Ref	Alt	Gene	Mutation type	Odds ratio	Case GT counts	Control GT counts	MAF_cases	MAF_controls	p-value
15	75913319	T	G	SNUPN	MISSENSE	2.939	0/910/4608	0/224/4162	0.0825	0.0255	2.718E-35
17	25973604	A	C	LGALS9	MISSENSE	2.868	0/908/4610	0/225/4161	0.0823	0.0256	1.43E-34
6	36979483	T	G	FGD2	MISSENSE	2.273	0/1101/4417	0/354/4032	0.0998	0.0404	1.648E-29
14	99976645	A	C	CCNK	SYNONYMOUS	2.317	0/992/4525	0/305/4080	0.0899	0.0348	7.488E-28
13	114188430	C	T	TMCO3	MISSENSE	2.117	0/1175/4327	0/402/3981	0.1068	0.0459	2.218E-27
14	99976639	G	C	CCNK	SYNONYMOUS	2.212	0/1028/4489	0/326/4059	0.0932	0.0372	3.666E-26
17	25973598	A	C	LGALS9	MISSENSE	2.884	0/521/4997	0/123/4263	0.0472	0.014	1.104E-21
14	77706020	A	C	TMEM63C	MISSENSE	2.491	0/637/4881	0/167/4219	0.0577	0.019	6.28E-21
11	117280516	A	C	CEP164	MISSENSE	2.118	0/826/4692	0/260/4126	0.0748	0.0296	2.482E-20
3	42739737	T	G	HHATL	MISSENSE	2.461	0/595/4923	0/160/4226	0.0539	0.0182	9.726E-20
3	48451952	A	C	PLXNB1	MISSENSE	2.364	0/620/4898	0/175/4211	0.0562	0.02	2.691E-19
12	56622883	A	C	NABP2	SYNONYMOUS	2.006	0/788/4730	0/264/4122	0.0714	0.0301	1.169E-17
2	85662149	A	C	SH2D6	MISSENSE	1.919	0/750/4765	0/271/4115	0.068	0.0309	4.265E-16
19	10946797	G	C	TMED1	MISSENSE	2.502	0/465/5053	0/112/4274	0.0421	0.0128	2.511E-15
14	105932775	G	C	MTA1	MISSENSE	1.986	0/692/4826	0/227/4159	0.0627	0.0259	3.137E-15
6	29429950	A	C	OR2H1	MISSENSE	2.459	0/444/5074	0/116/4270	0.0402	0.0132	3.615E-15
11	117280522	A	C	CEP164	MISSENSE	2.079	0/584/4934	0/174/4212	0.0529	0.0198	3.345E-14
2	85662154	A	C	SH2D6	MISSENSE	1.871	0/681/4835	0/253/4133	0.0617	0.0288	3.973E-14
13	88330245	A	C	SLITRK5	MISSENSE	3.142	0/306/5212	0/57/4329	0.0277	0.0065	1.126E-13
19	10946802	T	C	TMED1	SYNONYMOUS	2.33	0/404/5113	0/103/4283	0.0366	0.0117	3.074E-12
9	34564740	A	C	CNTFR	MISSENSE	1.913	0/569/4949	0/195/4191	0.0516	0.0222	3.569E-12
3	108474687	T	G	RETNLB	MISSENSE	2.392	0/362/5156	0/94/4292	0.0328	0.0107	5.77E-12
19	43025485	T	G	CEACAM1	MISSENSE	1.601	0/1016/4502	0/459/3927	0.0921	0.0523	7.691E-12
3	31659462	A	T	STT3B	MISSENSE	1.58	0/983/4534	0/451/3935	0.0891	0.0514	9.206E-12
12	109719316	T	G	FOXN4	MISSENSE	2.193	0/341/5177	0/101/4285	0.0309	0.0115	2.112E-10
8	145112936	T	C	OPLAH	SYNONYMOUS	2.159	0/365/5153	0/100/4286	0.0331	0.0114	5.016E-10
19	42799299	T	C	CIC	MISSENSE	3.234	0/196/4954	0/36/4111	0.019	0.0043	1.386E-09
6	41129252	C	T	TREM2	MISSENSE	4.225	1/100/5417	0/21/4365	0.0092	0.0024	1.401E-08
13	111164389	A	C	COL4A2	MISSENSE	1.633	0/571/4947	0/240/4146	0.0517	0.0274	1.611E-08
22	30951295	T	G	GAL3ST1	MISSENSE	2.595	0/225/5289	0/49/4324	0.0204	0.0056	1.804E-08

Table 2.3. SNPs reaching exome-wide significance ($p < 3.0 \times 10^{-7}$) under Models 1 and 2a after quality control. Model 2a included the covariates *APOE*, sex, sequencing center and PCs.

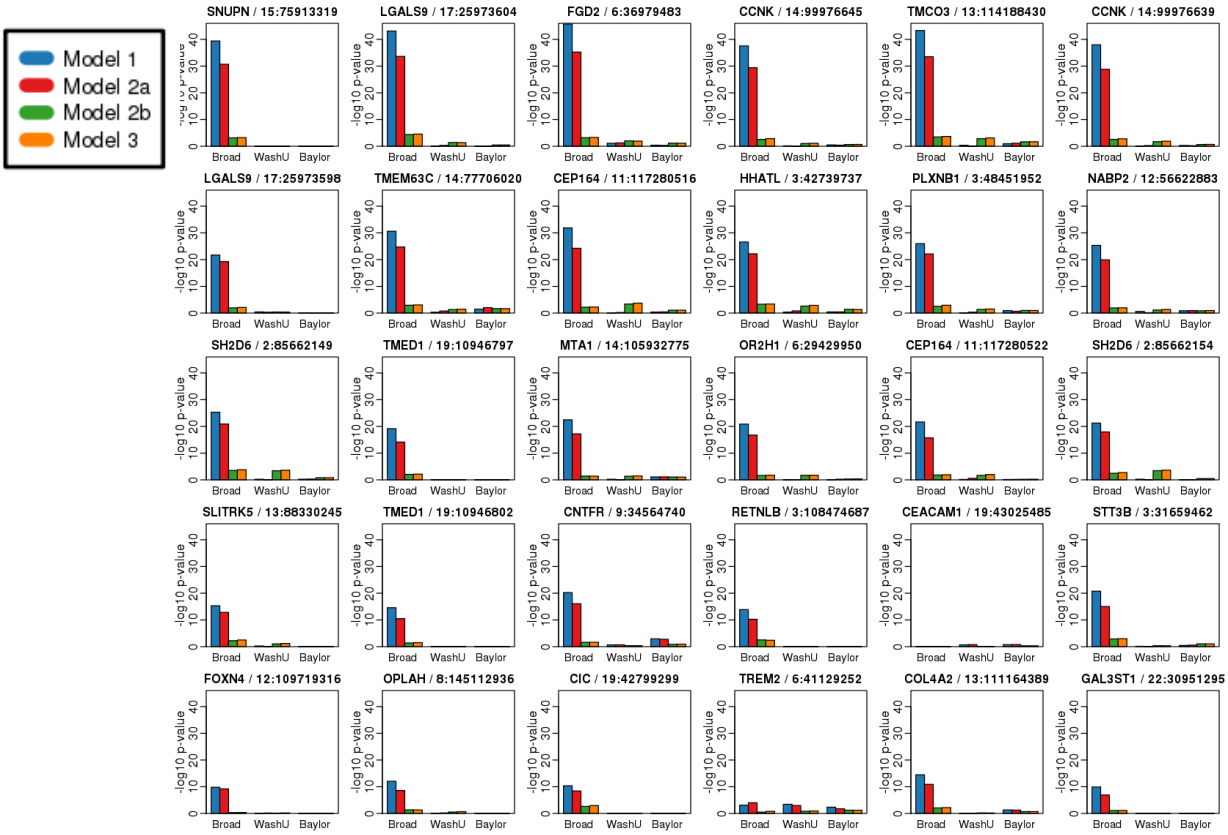

Figure 2.4. Log-transformed p-values for AD association under each model stratified by sequencing facility cohort. SNPs shown reached exome-wide significance under Models 1 and 2a in the full-dataset analysis.

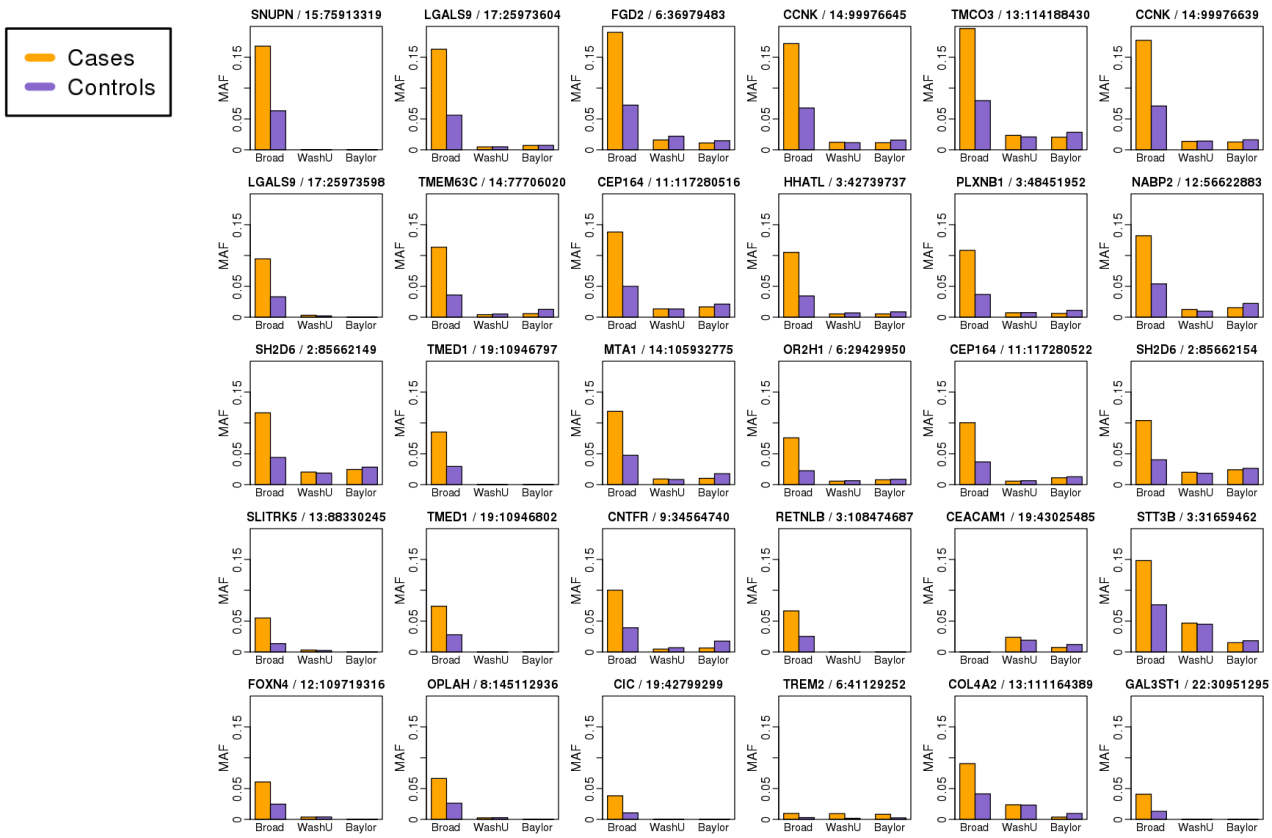

Figure 2.5. Minor allele frequencies in the cases and controls stratified by sequencing facility. SNPs shown reached exome-wide significance under Models 1 and 2a in the full-dataset analysis.

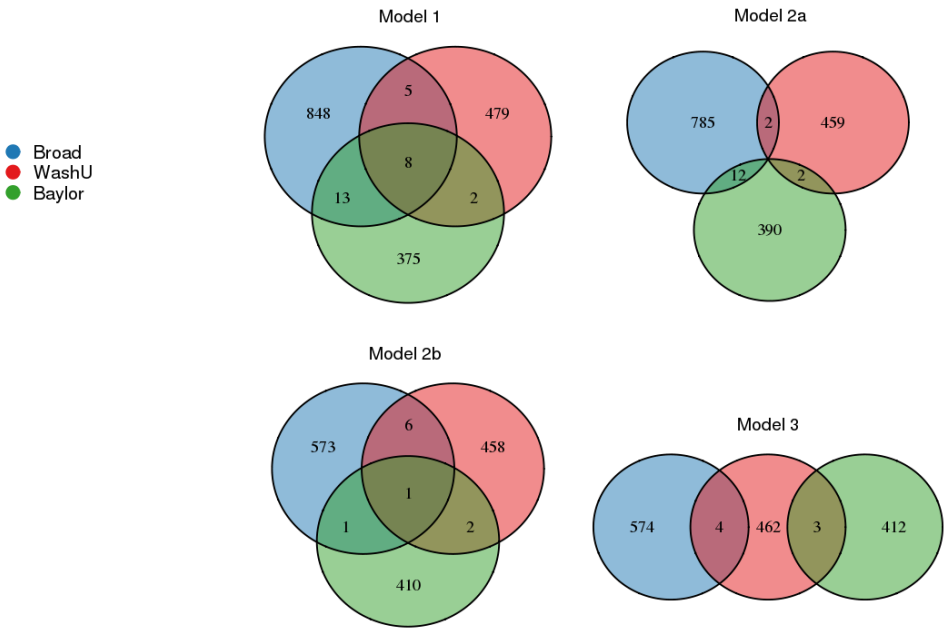

Figure 2.6. Number of nominally AD-associated ($p < 0.005$) QCed variants and their overlap among samples from three sequencing centers.

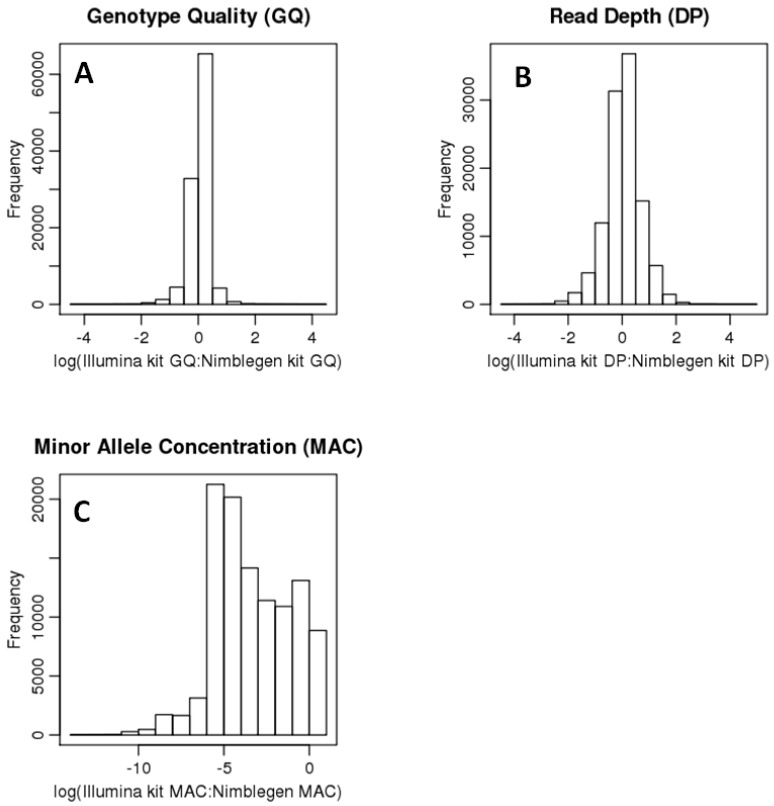

Figure 2.7. Global distributions of quality-metric ratios between QCed samples from each exome capture kit. Each data point represents the log of the ratio of the Illumina-captured samples' means to the Nimblegen-captured samples' means for a single variant.

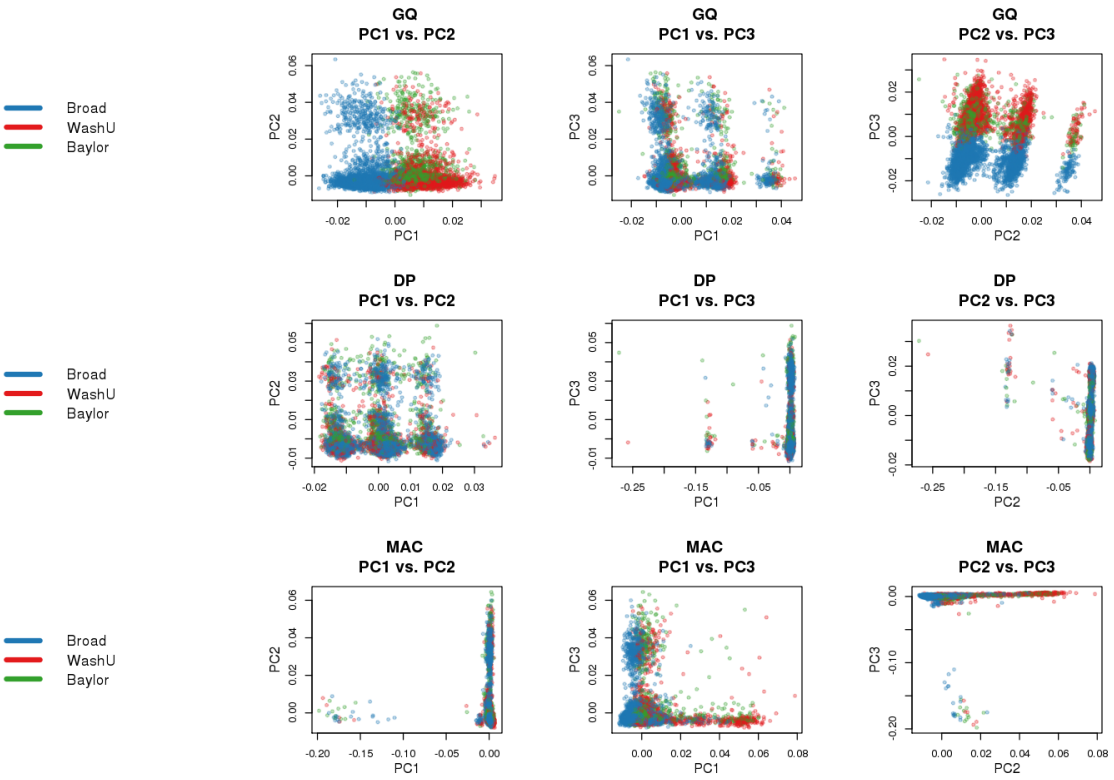

Figure 2.8. Principal component loadings at the 10% tails of quality-metric ratio distributions. Principal components were computed using as input the 9,904 samples' genotypes at the SNPs lying within the top 5% and bottom 5% of the distributions for GQ and DP, and the bottom 10% for MAC.

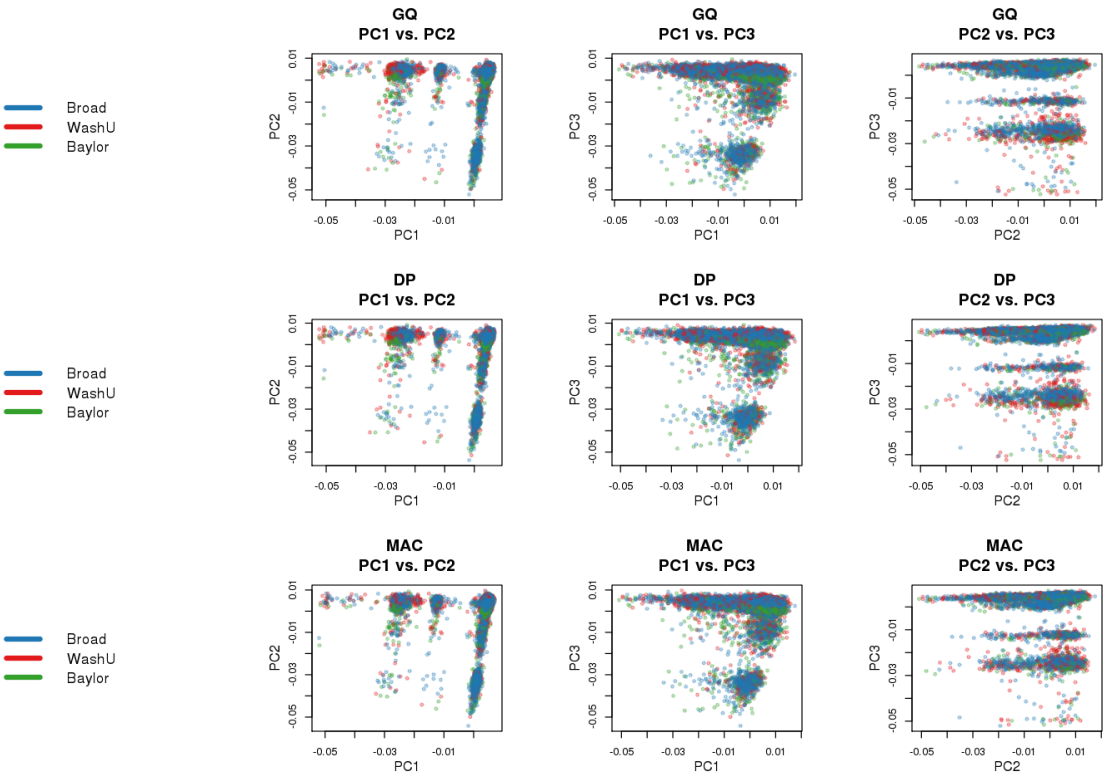

Figure 2.9. Principal component loadings at the middle 90% of quality-metric ratio distributions. Principal components were computed using as input the 9,904 samples' genotypes at the SNPs lying within the middle 90% of the distributions for GQ, DP and MAC.

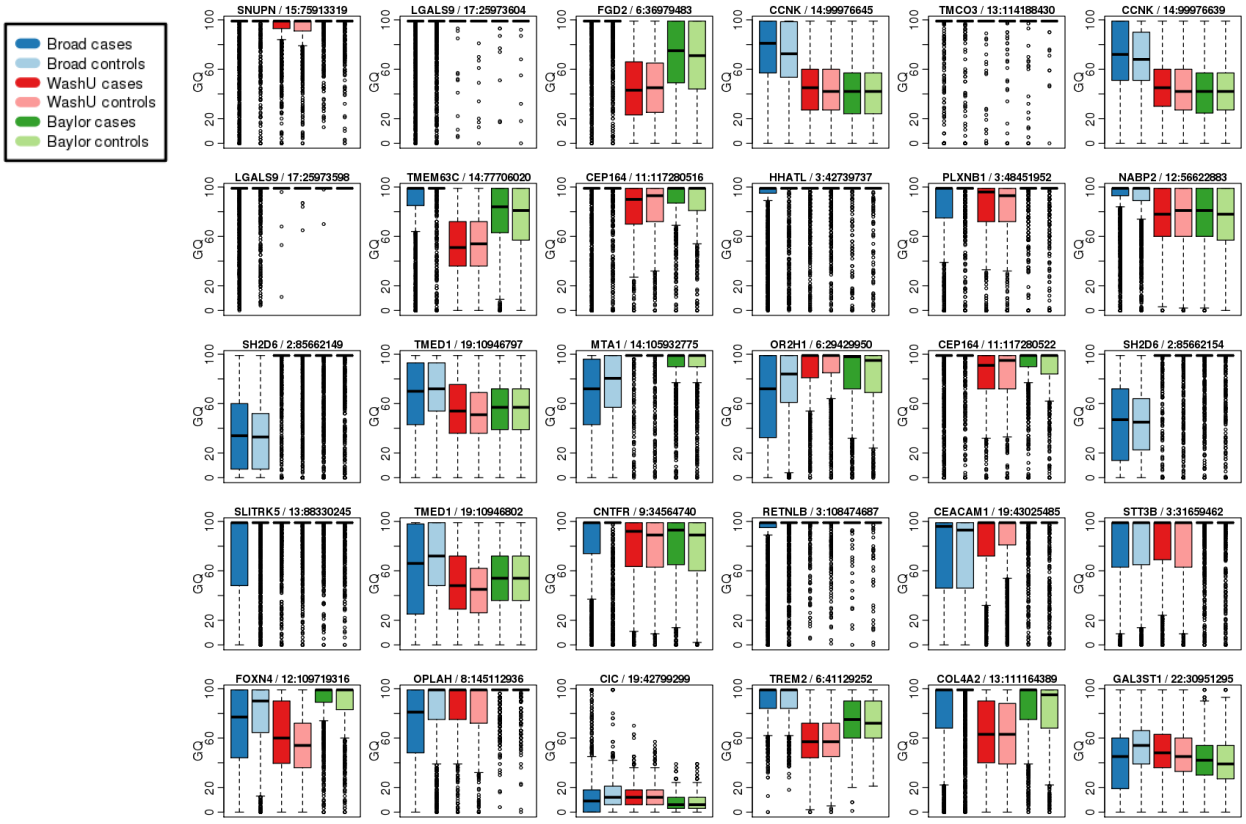

Figure 2.10. GQ distributions from the cases and controls stratified by sequencing center. SNPs shown reached exome-wide significance under Models 1 and 2a in the full-dataset analysis.

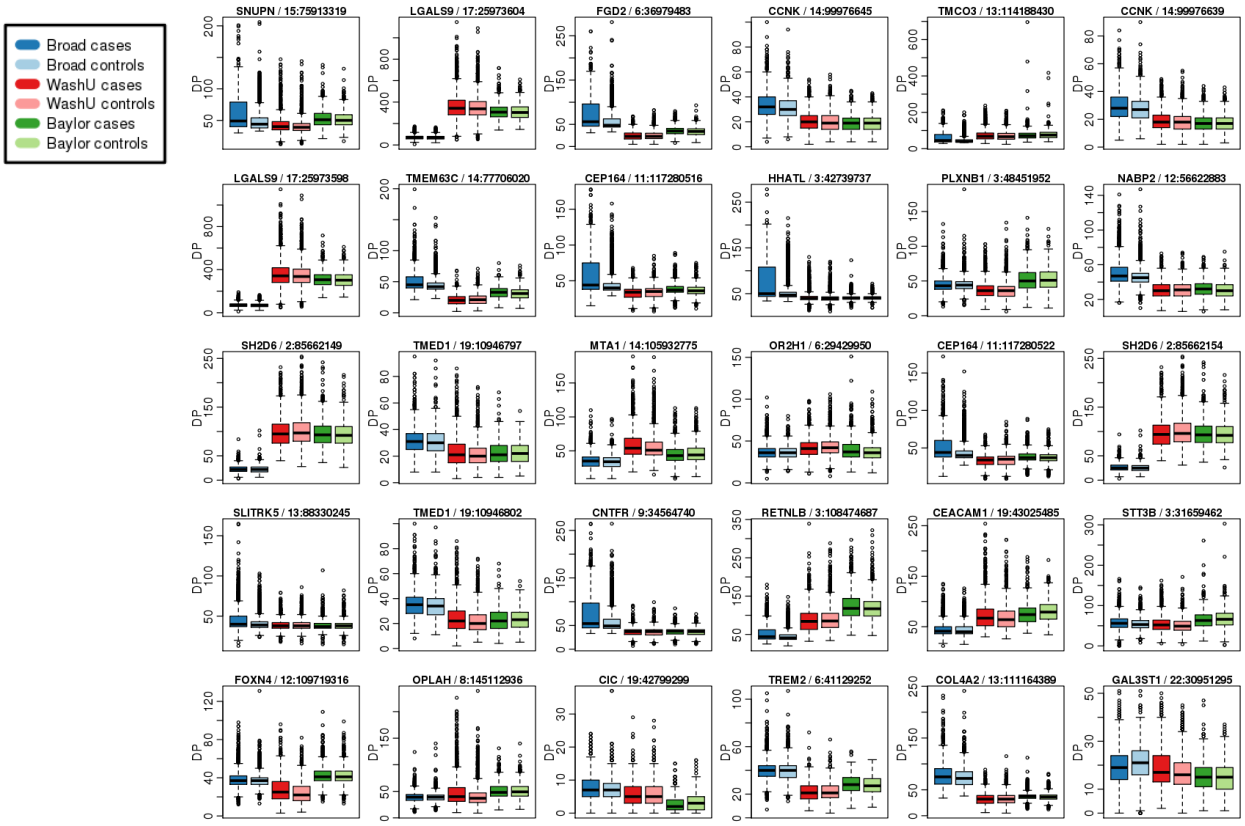

Figure 2.11. Read depth distributions from the cases and controls stratified by sequencing center. SNPs shown reached exome-wide significance under Models 1 and 2a in the full-dataset analysis.

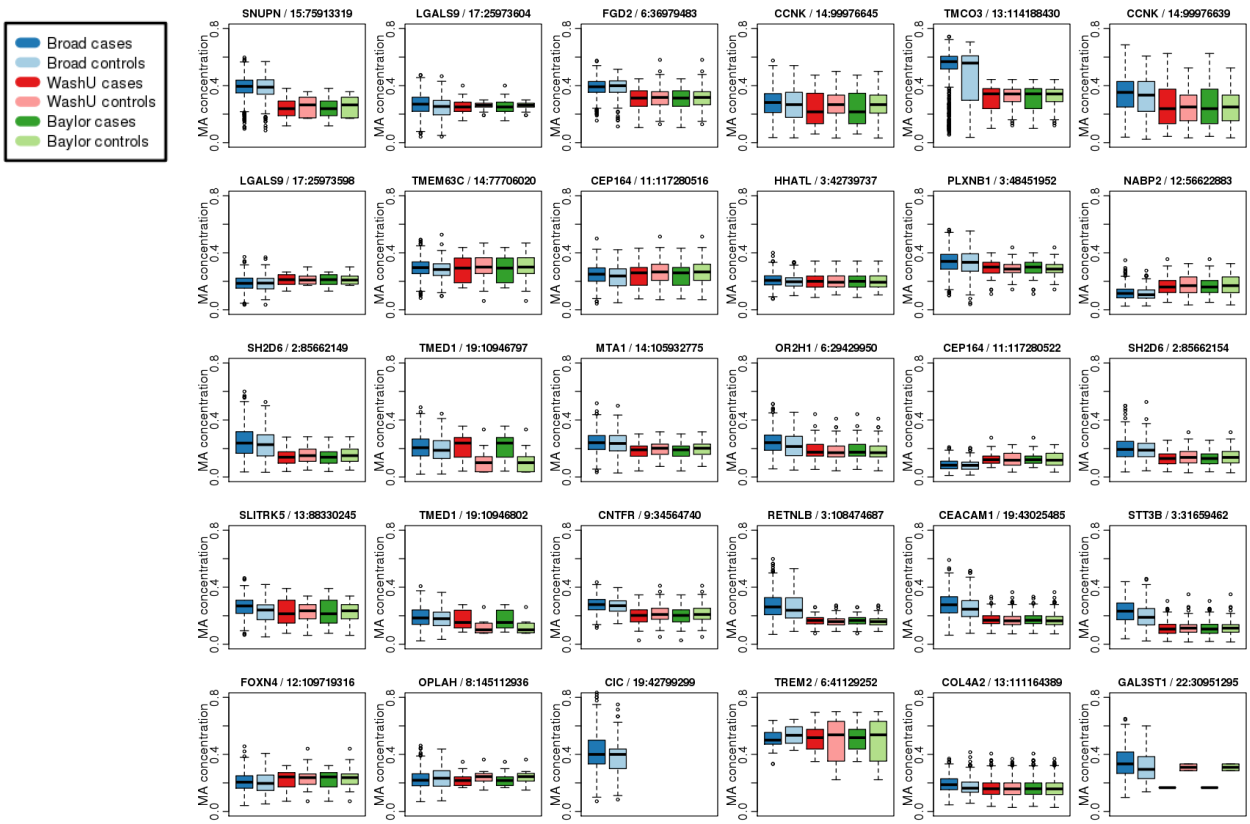

Figure 2.12. Minor allele concentration distributions from the cases and controls stratified by sequencing center. SNPs shown reached exome-wide significance under Models 1 and 2a in the full-dataset analysis.

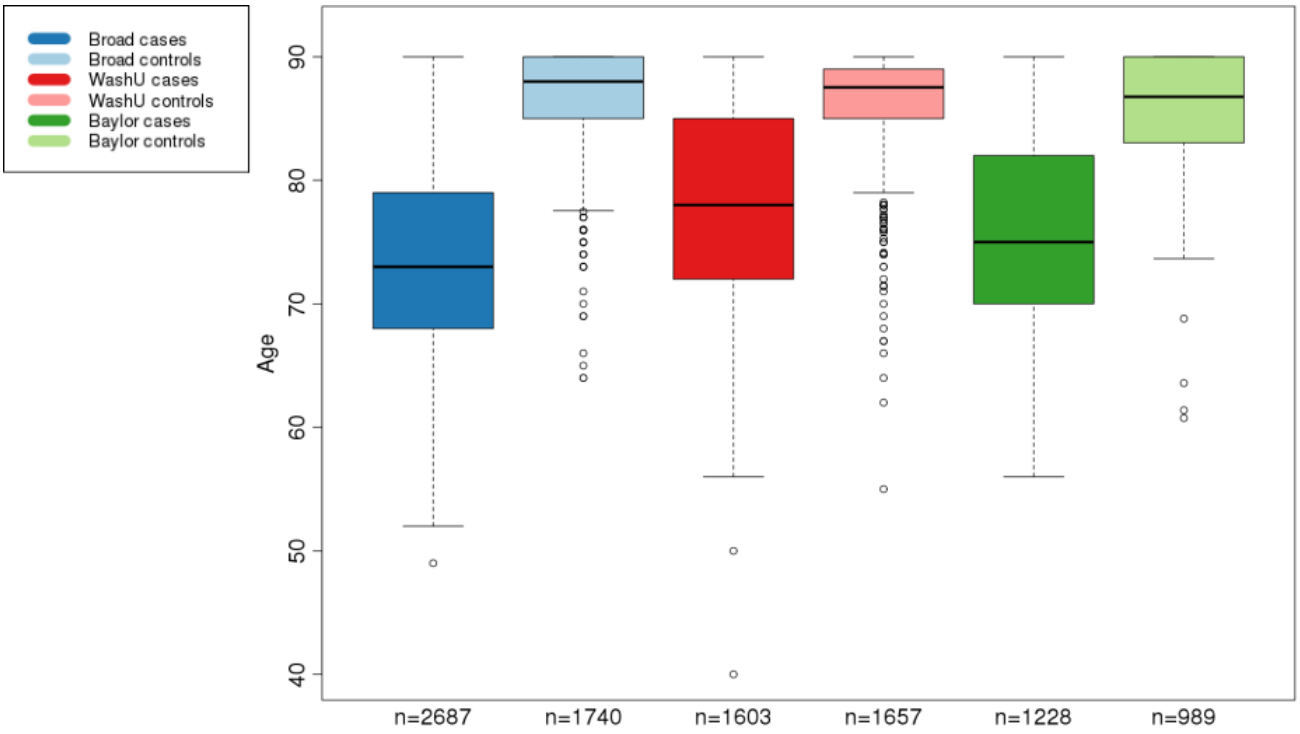

Figure 2.13. Age distributions for cases and controls from each cohort. The bold line inside each boxplot denotes the median. The lower and upper edges denote the first and third quartile, respectively.

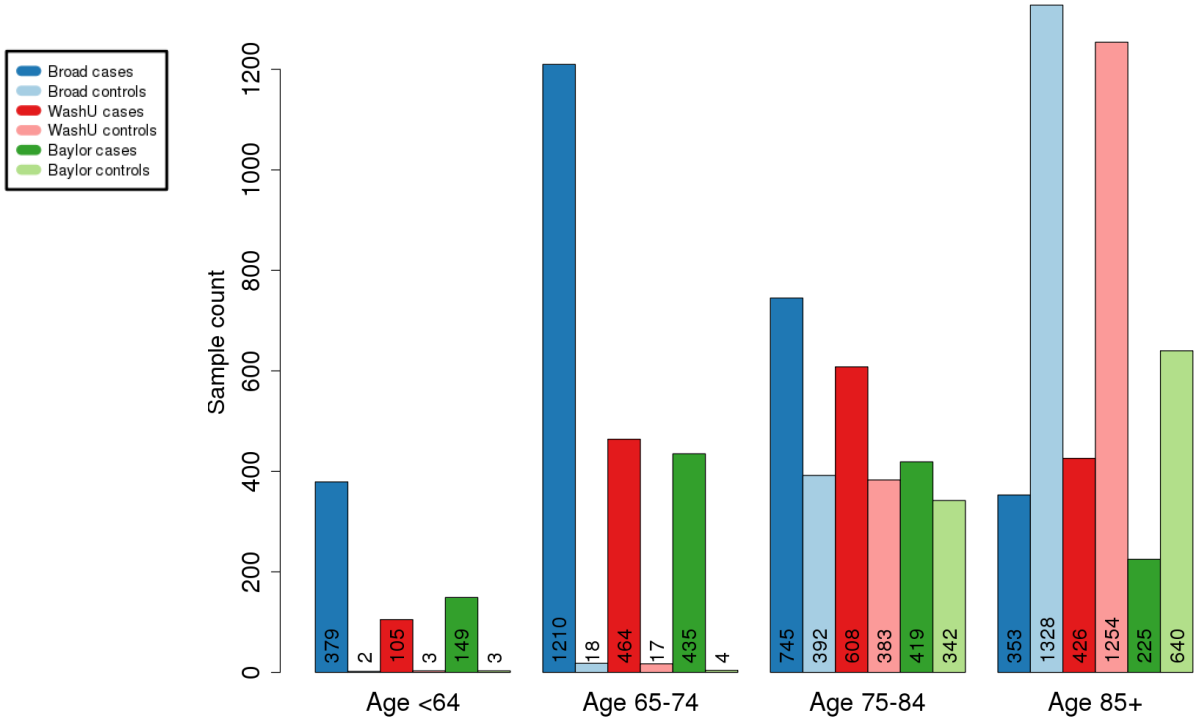

Figure 2.14. Number of case and control samples belonging to four age groups in each cohort.

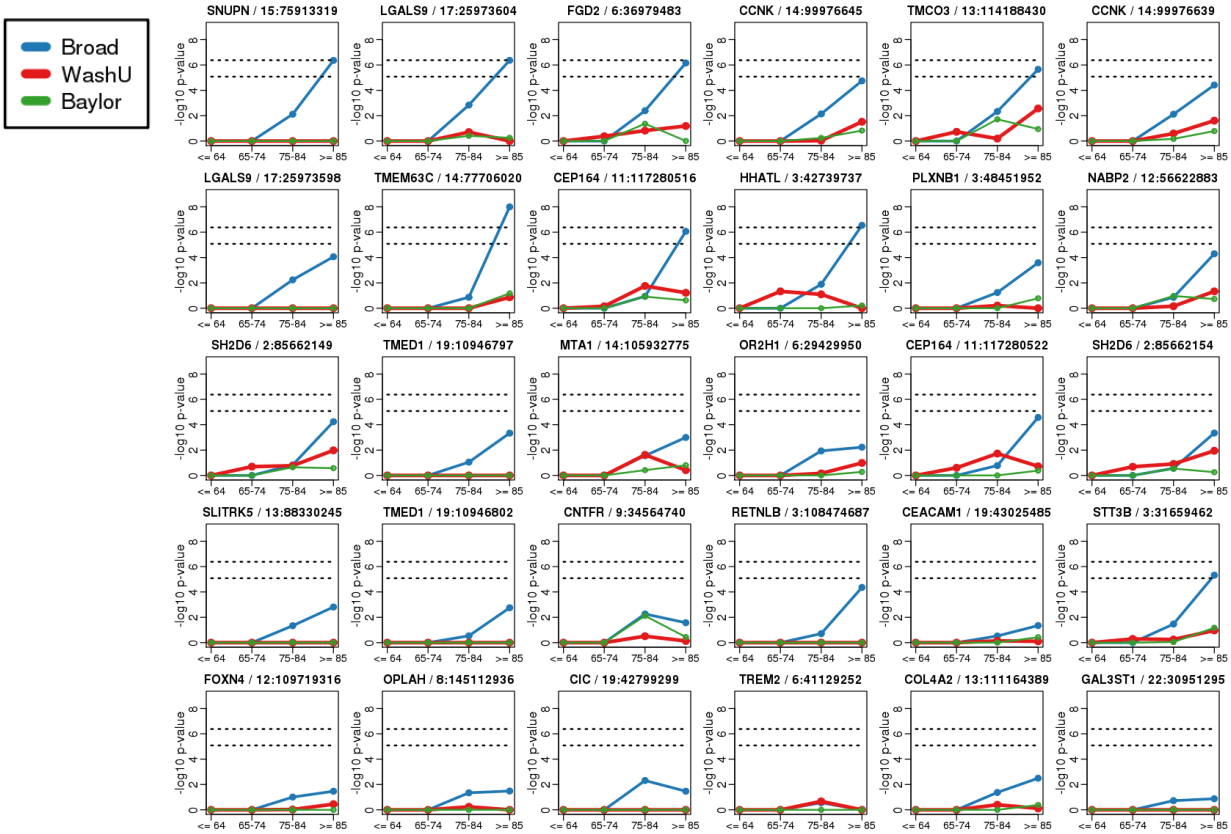

Figure 2.15. Log-transformed p-values for AD association stratified by sequencing facility and age group. The top dotted horizontal line denotes exome-wide significance in Broad and the bottom dotted horizontal line denotes suggestive significance in Broad. SNPs shown reached exome-wide significance under Models 1 and Model 2a in the full-dataset analysis.

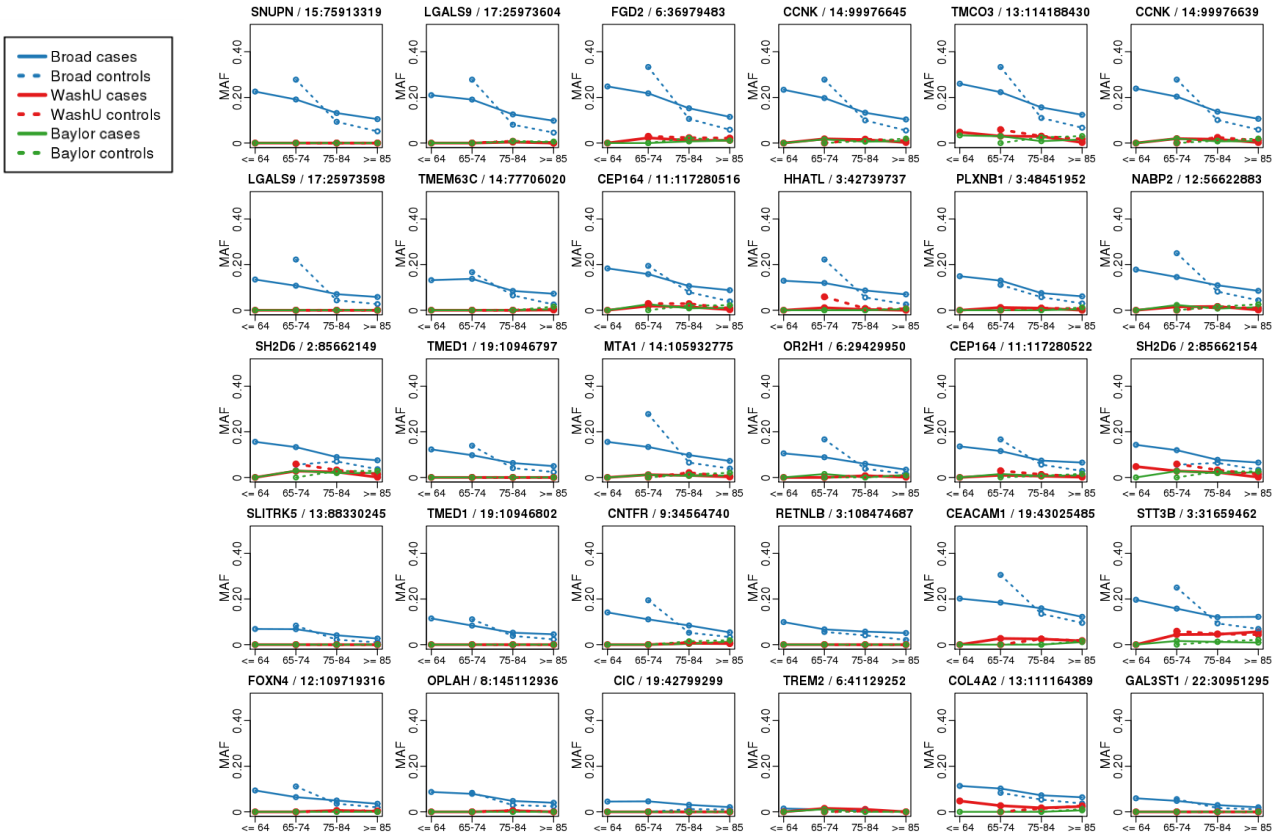

Figure 2.16. Minor allele frequencies in cases and controls stratified by sequencing facility and age group. SNPs shown reached exome-wide significance under Models 1 and 2a in the full-dataset analysis. Allele frequencies of controls in the youngest group are omitted because each center had fewer than 5 individuals represented.

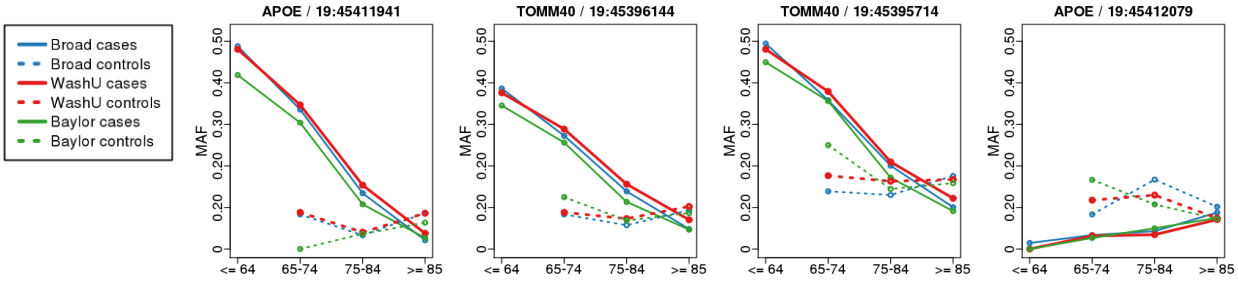

Figure 2.17. Minor allele frequencies in cases and controls stratified by sequencing facility and age group. SNPs shown were shared among all three sequencing facilities and reached genome-wide significance under Model 1 in the full-dataset analysis. Allele frequencies of controls in the youngest group are omitted because each center had fewer than 5 individuals represented.

CHAPTER 3: A COMPARISON OF GENOTYPING-BY-SEQUENCING (GBS) ANALYSIS METHODS ON LOW-COVERAGE CROP DATA SHOWS ADVANTAGES OF A NEW WORKFLOW, GB-EASY

BACKGROUND

Genotyping-by-sequencing (GBS), a simplified reduced-representation sequencing approach (Elshire et al. 2011), has gained popularity in crop research and plant breeding for high throughput, low-cost genotyping. It has been applied to projects ranging from genomic selection to gene mapping to genome-wide association studies in numerous crop species (Furuta et al. 2017; Liu et al. 2014; Poland et al. 2012; Sonah et al. 2015; Wu et al. 2016). GBS relies on restriction enzymes to generate a reduced representation of the genome for sequencing. Compared to other approaches such as RAD-seq, the GBS library preparation protocol involves fewer steps, requires less DNA, and lacks a size selection step (Elshire et al. 2011). In GBS, DNA samples are digested and ligated to barcoded adapters in single wells, pooled, and then enriched by PCR. In contrast to the relatively simple and straightforward library preparation, GBS data analysis is complicated by the nature of the random location, reduced-representation approach.

Bioinformatics software packages and workflows have been developed to provide the architecture for analysis of reduced-representation sequencing data (Catchen et al. 2013; Sonah et al. 2013; Torkamaneh et al. 2017). Several of these platforms utilize the same tools and algorithms commonly applied to whole-genome sequence data, while others utilize algorithms developed specifically for GBS and RAD sequencing. Although designed to facilitate and simplify data processing, these GBS pipelines nevertheless can be difficult for non-specialist researchers such as plant breeders to install or implement. Issues include high levels of complexity, requirements for additional libraries or uncommon packages, or additional processing steps outside of the pipelines. A different approach, TASSEL / TASSEL-GBS (Bradbury et al. 2007; Glaubitz et al. 2014), provides an all-in-one desktop software package that is easy to install and use, and performs both GBS data processing and genetic analysis using the resources

Parts of this chapter were published as Wickland et al, 2017: *A comparison of genotyping-by-sequencing analysis methods on low-coverage crop datasets shows advantages of a new workflow, GB-eaSy*. BMC Bioinformatics 18(1): 586.

of a stand-alone PC. However, while this software is widely adopted in cereal genetics, it was optimized for use in maize, and uses heuristics such as the reduction of reads to tags before alignment to enable reasonable run times on PC hardware. These heuristics are less clearly advantageous in recently polyploid species; for this reason, others (e.g. Torkamaneh et al. 2017) have developed different approaches for crops such as soybean. Finally, the all-in-one software package approach means that users cannot themselves easily modify TASSEL-GBS to accommodate new sequencing technology or other software packages.

More recently, known segregating sites from pan-genome data have been shown to substantially improve accuracy and yield from reduced-representation sequencing (Lu et al. 2015); however, for other crops such as soybean and many others important for food production, population-level diversity is not yet sufficiently well characterized at the whole-genome level, and better tools to identify SNPs *ab initio* are still needed. In addition, recently polyploid genomes such as soybean (Schmutz et al. 2010) present a complication to the performance of alignment and variant calling for all forms of reduced-representation sequencing. This may influence the performance of different approaches relative to more straightforward diploid genomes.

Here we present GB-eaSy, a GBS bioinformatics pipeline that efficiently incorporates widely used genomics tools, parallelization and automation to increase the accuracy, efficiency and accessibility of GBS analysis. GB-eaSy has been specifically developed to be straightforward to install and use on typical UNIX / HPC hardware, to contain readily updateable public software where possible, and to match or exceed the performance of current GBS SNP-calling methods used on soybean or other complex, repetitive and recently polyploid genomes. It can process reduced-representation data from any organism with a reference genome. We compared the performance of GB-eaSy to four other GBS bioinformatics data analysis platforms using low-coverage Illumina sequence data from three soybean populations. GB-eaSy rapidly and accurately identified the greatest number of SNPs across all three populations, with SNP calls in close agreement with whole-genome sequencing of selected lines. In addition, the unexpectedly low convergence between the five analysis methods but generally high accuracy indicated that the workflows arrived at largely complementary sets of valid variant calls.

MATERIALS AND METHODS

Samples

GBS libraries were constructed from three soybean populations (**Table 3.1**). Population 1 consisted of 378 F2 lines resulting from a cross between the accession Prize and an NMU-mutagenized individual from the reference genotype Williams 82. Population 2 contained 391 F2 individuals from a cross between two breeding lines. Finally, Population 3 consisted of 81 unrelated accessions (with 2-4 replications) that form an association panel. DNA was extracted using the CTAB method (Richards et al. 2001) except for the Prize x NMU-mutagenized Williams 82 population (Population 1), which used the E-Z 96 Plant DNA kit (Omega Bio-Tek, Norcross, GA). All libraries were sequenced at low coverage typical of plant breeding experiments, with coverage varying from 1.87x to 4.47x.

GBS library preparation

GBS libraries were prepared according to the two-enzyme protocol described in Poland et al. 2012 with minor modifications (kindly provided by Dr. P. Brown, UC Davis). Two-enzyme pairs (HindIII-MseI and HindIII-BfaI) were used to achieve a balanced representation of HindIII cut sites. In brief, restriction and ligation were carried out simultaneously, followed by PCR amplification. First, 5 µl of DNA (25-50 ng/µl, 125-250ng total) from each sample was pipetted into its own well on a 384-well plate that contained restriction-ligation master mix. The master mix in each well consisted of 2.5µl 10X NEB CutSmart buffer (final concentration 1X), 2.5µl 10 mM dATP (final concentration 1 mM), 0.1µl (2U) HindIII, 0.2µl MseI or BfaI, 0.1µl concentrated T4 DNA ligase (40U), 0.5µl each of 10uM adapters, and 14.1 µl molecular biology-grade water. The barcoded “rare adapters” were designed to anneal to the cut HindIII site, while the non-barcoded “common adapters” annealed to the cut MseI or BfaI site.

The 384-well plates underwent digestion and ligation in the thermocycler at 37°C for 1 min, 25°C for 1 min, repeated 100 times. Next, 8 µl from each well was pooled into a 1.5mL microfuge tube, cleaned using Agencourt AMPure XP beads (Beckman Coulter Life Sciences, Indianapolis, Indiana, USA), dried, and suspended for PCR amplification in a solution of Phusion

Master Mix (NEB, Ipswich, MA). PCR settings for amplification were 98°C for 30s, 15 cycles (98°C for 10s, 68°C for 30s, 72°C for 30s), 72°C for 5m, followed by 4°C until sample recovery. Next, AMPure cleanup was repeated, and the resulting library was evaluated on a Bioanalyzer 2100 (Agilent, Santa Clara, CA) using a DNA7500 chip to assess amplification success, fragment size, and DNA concentration. Finally, each library was diluted to 10 nM DNA in LIB buffer (10mM Tris-HCL (EB) w/ 0.05% Tween-20) and run on either an Illumina HiSeq2500 or HiSeq4000 using the HiSeq SBS sequencing kit version 4 at the Roy J. Carver Biotechnology Center at the University of Illinois at Urbana-Champaign.

GBS data analysis platforms

TASSEL-GBS

TASSEL-GBS was developed to assign SNP genotypes from GBS data in a time- and storage-efficient manner (Glaubitz et al. 2014) (**Table 3.2**). Unlike SNP calling for whole-genome data, which involves first aligning all reads to the reference genome and then calling SNPs, TASSEL-GBS dramatically reduces computational demands by consolidating reads into a master "tag list" containing the unique sequences. This tag list is then aligned to a reference genome. For species lacking a reference genome, the consensus allele at each position is considered the reference allele. Variant identification in the TASSEL5GBSv2 pipeline (https://bitbucket.org/tasseladmin/tassel-5-source/wiki/Tassel5GBSv2Pipeline) consists of two main steps: SNP discovery and production SNP calling. In SNP discovery, TASSEL-GBS determines SNPs and SNP coverage within each tag for each sample and outputs the results to a database. In production SNP calling, SNP genotypes in each sample are output. Each step is performed internally with TASSEL-GBS plugins, except alignment, which is carried out externally using software such as BWA-MEM (H. Li 2013). Prior to running TASSEL, we removed adapter sequence from the reads using cutadapt (Martin 2011) after finding that adapter contamination severely impaired the accuracy of TASSEL-GBS SNP calls relative to the other methods.

Stacks

Stacks is a software package developed for RAD sequencing that identifies SNPs and calculates population statistics from any restriction enzyme-based, reduced-representation sequence data (Catchen et al. 2013) (**Table 3.2**). After demultiplexing and cleaning the sequenced reads, Stacks assembles loci from each sample (with or without a reference genome) and groups together loci across samples to construct a catalog. Comparison between the catalog and loci from each sample allows inference of SNPs and genotypes. Optional additional steps include creation of genetic maps and calculation of population statistics. Like TASSEL-GBS, each step except alignment (here performed by BWA-MEM) uses the software's internal algorithms.

IGST

IGST (IBIS Genotyping by Sequencing Tools) processes GBS data by implementing several popular genomic software tools connected by Perl and Python scripts (Sonah et al. 2013) (**Table 3.2**). After setting up a predefined directory structure and naming input files according to a specific convention, the user issues a single command that runs the entire pipeline. IGST demultiplexes and cleans barcoded reads using Sabre (https://github.com/najoshi/sabre), aligns demultiplexed reads to the reference genome using BWA-ALN (H. Li and Durbin 2009), converts the aligned sequences to BAM format using SAMtools (H. Li 2011), and identifies SNPs using SAMtools and BCFtools (H. Li 2011). The resulting SNP calls are filtered by VCFtools (Danecek et al. 2011).

Fast-GBS

Fast-GBS follows a strategy similar to IGST but employs a different alignment algorithm, a different variant caller, and a bash script that runs each software program (Torkamaneh et al. 2017) (**Table 3.2**). As with IGST, the user must set up a predefined directory structure and name files according to a specific convention before inputting a single command to run the workflow. This pipeline demultiplexes reads using Sabre, trims and cleans reads using Cutadapt, aligns reads to the reference genome using BWA-MEM, and calls variants using Platypus (Rimmer et al. 2014). As a haplotype-based

variant caller, Platypus identifies single-allele SNPs as well as compound SNPs consisting of short strings of adjacent alleles. To facilitate comparisons with the other pipelines, we used the VariantsToAllelicPrimitives script within the Genome Analysis Toolkit (Van der Auwera et al. 2013) to deconvolute the multi-allelic SNPs into individual allelic primitives, as recommended by Baes et al. 2014.

GB-eaSy

The GB-eaSy pipeline developed for this project consists of a Bash shell script that executes several bioinformatics software programs in a parallel UNIX / Linux environment. This workflow requires a reference genome and is compatible with both single- and paired-end Illumina reads. Its name derives from its straightforward, transparent implementation of GBS variant calling; GB-eaSy is appropriate for users without extensive command-line expertise as well as for experienced bioinformaticians who may choose to modify any step of the script. GB-eaSy implements the same well-tested and regularly updated tools commonly adopted in whole-genome sequencing. In contrast to some GBS pipelines, GB-eaSy does not require the user to follow strict instructions regarding directory structure or file names; instead, the Bash script performs these steps automatically. The GB-eaSy shell script, a walkthrough of each command, and a tutorial using sample data are hosted at https://github.com/dpwickland/GB-eaSy.

Before starting the pipeline, the user modifies a parameters file with settings customized for their GBS project (e.g. path to raw sequencer output file, path to barcodes file, number of CPU cores to use). The user then issues a single command to execute the pipeline. The first step of GB-eaSy uses the software GBSX (Herten et al. 2015) to demultiplex reads and trim adapter sequences based on a user-created barcodes file containing the short barcode sequences that uniquely identify each sample; for our study, we modified the GBSX script (GBSX.jar) to include the HindIII cut site, which was not supported initially. Next, demultiplexed reads are aligned to the reference genome using BWA-MEM; GB-eaSy hastens this alignment step by processing

read files in parallel using GNU Parallel (Tange 2011). After alignment, BCFtools is used to create a pileup of read bases from which it calls SNPs. This SNP-calling step uses GNU Parallel to process each entry in the reference genome file (e.g. each chromosome, each scaffold) on its own CPU core, greatly increasing the efficiency of SNP identification. Finally, the output VCF file is filtered by VCFtools according to a user-specified minimum read depth (**Table 3.2**).

Whole-genome sequencing

To validate the output from the GBS pipelines, Illumina whole-genome sequence (WGS) data was obtained (experimentally in the case of Prize for Population 1 and the case of LG12 for Population 2, or from the data obtained by Song et al. 2017 for four lines of the soybean NAM association panel for Population 3) for comparison of GBS and WGS SNP calls (**Table 3.3**). As with the GBS pipelines, WGS reads were aligned to the reference genome using the software BWA-MEM. However, variant calling on the WGS datasets was carried out with GATK HaplotypeCaller (Van der Auwera et al. 2013), a tool not used by any of the GBS pipelines, to provide independent assessment of GBS SNP call accuracy.

Pipeline comparisons

The five GBS pipelines and the WGS pipeline described above were run with the following parameters to make the analysis as equivalent as possible between workflows: minimum read length of 80 bases after adapter and barcode trimming, minimum base quality of 20 and minimum mapping quality of 20 for variant calling (corresponding to a 1 in 100 chance of an incorrect base call or mapping call, respectively), and identification of SNPs only (no indels). Other parameters were set at default values. The software package VCFtools was then used to remove SNP calls supported by less than 2 reads (i.e. minimum depth of 2 reads) to increase the reliability of distinguishing homozygous from heterozygous genotypes (note that our lowest coverage dataset has an average depth per sequenced base of 1.87x). Recent versions* of component software packages and commands were used for each pipeline, with the following exceptions: for IGST, commands were run using SAMtools version 0.1.18 and Picard version 1.119 because the IGST workflow was incompatible with later versions. Finally,

11 CPU cores were used at any steps that offered an option for parallelization. In-house scripts, BCFtools and VCFtools were used to compute and compare the number of chromosomal SNPs identified by the pipelines and to calculate missing data values. All programs were run on a Linux server with two Intel® Xeon® X5650 processor chips, each with six CPU cores, and 48 GB RAM.

* BWA 0.7.15-r1140
Picard 2.10.0
VCFtools 0.13
SAMtools/BCFtools 1.5
JAVA 1.8.0_121
GBSX_v1.3
GNU parallel 20170122
Cutadapt 1.12
TASSEL 5.0, build April 6, 2017
Stacks 1.46
Platypus 0.8.1

RESULTS

GBS SNP calls and their agreement with WGS SNP calls

We compared the SNP calls within and between pipelines on three different populations. Populations 1 and 2 were each 384-well plates used to sequence populations of F2 individuals chosen to mimic mapping populations or breeding studies, while Population 3 was a set of 81 diverse lines, again replicated across a 384 well plate, that can be used as a GWAS diversity panel (Song et al. 2017). Population 1 was derived from a cross between Prize (a US-adapted cultivar) and Williams 82 (the target of the reference genome project (Schmutz et al. 2010), while Population 2 was derived from a cross between two breeding lines that should be equally distant from the reference genome. After preparing GBS libraries and obtaining low-coverage Illumina sequence data (ranging from 1.87 to 4.47x depth per sequenced base), we called SNPs using the five pipelines and computed the total number of SNPs identified and the number of SNPs shared between pipelines. In addition, we compared the GBS SNP calls to WGS SNP calls of selected lines to calculate the SNP concordance and allelic concordance between

GBS and WGS. The analysis excluded indels to simplify comparisons among the methods because some methods call only SNPs and because SNPs are the markers of choice in most breeding projects. All SNPs were called relative to the Williams 82 soybean reference genome.

In terms of SNP yield, the relative ranking of each pipeline remained similar across all three populations: GB-eaSy called the most SNPs, followed in order by Fast-GBS, IGST and Stacks (rank depending on population), and TASSEL-GBS (**Figure 3.1**). In Population 1, the number of SNPs identified ranged from 35,328 (TASSEL-GBS) to 88,298 (GB-eaSy). Population 2 had the greatest number of SNP calls, ranging from 88,423 (TASSEL-GBS) to 249,472 (GB-eaSy); the comparatively large SNP yield of Population 2 likely resulted from the HiSeq4000 outputting 150,000 more reads than the HiSeq2500 used with Populations 1 and 3 (**Table 3.1**). In Population 3, the number of SNPs called ranged from 78,848 (TASSEL-GBS) to 163,571 (GB-eaSy). Within each population, a small portion of SNPs was called by all five workflows, with the proportion of convergent SNPs being roughly consistent (**Figure 3.2A**). A similar trend appears in the data for individual soybean lines (**Figure 3.2B**).

Because the SNP concordance between GBS analysis platforms was unexpectedly low (**Figure 3.2**), whole-genome data of six lines was obtained for comparison of GBS and WGS SNP calls. To avoid biasing these comparisons in favor of a particular GBS platform, GATK HaplotypeCaller (a tool not used by any of the GBS workflows) was used to call SNPs in the WGS datasets. The GBS data for these individual lines follows the population-level pattern of GB-eaSy finding the most GBS SNPs, closely followed by Fast-GBS (**Figure 3.3A**). SNP concordance was calculated as the percentage of GBS SNP sites (e.g. chromosome 1, position 8144) that were also identified by WGS (**Figure 3.3B**). Depending on the line under study, either Stacks, TASSEL-GBS or IGST exhibited the highest SNP concordance with WGS. Across all pipelines, SNP concordance was relatively lower in the lines Magellan, Maverick, Prohio and Skylla due to the low coverage of their WGS data (ranging from 2.02x to 5.37x) and therefore fewer sites sampled (**Figure 3.3B**).

We also assessed the allelic agreement (e.g. chromosome 1, position 8144, nucleotide C) between GBS SNP calls and WGS SNP calls for the set of concordant SNPs identified above (**Figure 3.3C**). In every line examined, GB-eaSy, TASSEL-GBS and IGST all achieved high allelic agreement (above 99%) with WGS, Fast-GBS reached allelic agreement between 97.19% and

99.54%, and Stacks reached allelic agreement between 95.55% and 98.45%. While GB-eaSy, TASSEL-GBS and IGST attained similarly high WGS-agreement rates, GB-eaSy identified the greatest number of SNPs in allelic agreement with WGS in each line (**Figure 3.3D**).

Missing data

GBS, unlike RAD-seq used for biological diversity analysis, is tuned to identify as many SNPs as possible, with missing data accounted for in later analysis by imputation of haplotypes using reference genome data. However, any GBS data analysis must consider the large proportion of missing/unsampled data, which can often be a limiting factor in downstream applications of the genotype data. The more sensitive a method is to polymorphisms with lower coverage, the more missing data in percentage terms is likely to be observed when comparing samples; therefore, the key parameter is the outright number of SNPs that are present in a sufficient proportion of lines for the analysis to be used. Within the three populations, the average percentage of sampled SNPs not present in any given line was fairly consistent: 83.4% (GB-eaSy) to 89.7% (Stacks) in Population 1, 59.4% (TASSEL-GBS) to 71.5% (GB-eaSy) in Population 2, and 62.4% (TASSEL-GBS) to 69.6% (GB-eaSy) in Population 3 (**Table 3.4**). In Population 1, GB-eaSy found the most SNPs present in at least 25% and 50% of sampled lines, while TASSEL-GBS found more SNPs present in at least 75% and 90% of sampled lines (**Table 3.4**). In Population 2, Stacks identified the most SNPs present in at least 25% of lines, GB-eaSy identified the most present in at least 50% and 75% of lines, and TASSEL-GBS identified the most SNPs in at least 90% of lines. Finally, in Population 3, Fast-GBS found the greatest number of SNPs present in at least 25% of lines, while GB-eaSy found the greatest number of SNPs present in at least 50%, 75% and 90% of lines. In this case, the variation in performance across the three populations was substantial, but GB-eaSy showed the best or among the best performance for each population. Notably, since each pipeline produces a different subset of valid SNPs (**Figure 3.2B**), the optimal strategy for minimizing missing data is likely the combination of multiple approaches.

Run time and disk space

The pipelines differed widely in their time to completion. TASSEL-GBS (including the initial Cutadapt step) finished most rapidly for each population (**Table 3.5**), as expected from its extensive use of tag heuristics to speed alignment. Fast-GBS and GB-eaSy alternately ranked as second and third fastest, depending on the population and the total number of reads. Stacks and IGST used the most wall-clock time per sample, with IGST taking at least three times as long as TASSEL-GBS in every population.

The disk space required paralleled the run time in most pipelines (**Table 3.6**). For each population, TASSEL-GBS required the least amount of storage. GB-eaSy and Stacks used approximately twice the disk space required by TASSEL-GBS. Despite their parameters being set to delete intermediate files where applicable, IGST and Fast-GBS used substantially more disk space than the other methods.

DISCUSSION

Despite the availability of multiple tools for GBS data processing, a need exists for a GBS pipeline that is easy to install, interfaces with standard tools, is optimized for high density SNP calling in polyploid crop genomes, and quickly and reliably identifies a large number of accurate SNPs while minimizing its storage footprint. We developed GB-eaSy, a GBS bioinformatics pipeline suitable for both command line novices and experienced bioinformaticians, and aim it primarily at the soybean community, where use of such processing software is increasing. However, GB-eaSy should be applicable to any non-model plant species with a reference genome, particularly to polyploids with repetitive genomes such a soybean. The 1.1-gigabase, recently paleopolyploid soybean genome contains multiple copies of 75% of its genes (Schmutz et al. 2010), which presents challenges to accurate processing of genomic data. Therefore, soybean qualifies as a suitable test subject to assess the accuracy of GB-eaSy's SNP calls. Comparison of GB-eaSy to other GBS data workflows indicated that GB-eaSy rapidly and accurately identified the most SNPs in all three soybean populations examined, without demanding excessive disk space.

Different SNP calling strategies

A key difference among GBS pipelines that may explain their discrepant results is the software used for variant calling, and its approach to determining the consensus genotype in a group of reads and whether that consensus varies from the reference. Both IGST and GB-eaSy use BCFtools/SAMtools as the variant caller, which relies on a Bayesian strategy to select as the consensus genotype at a given locus the base with the highest Phred score that maximizes the posterior probability (O'Rawe et al. 2013). If the consensus genotype at the locus differs from the reference, a SNP is called. Previous work has validated the accuracy of the BWA and SAMtools/BCFtools combination used in IGST and GB-eaSy. For instance, Hwang et al. 2016 evaluated thirteen variant calling pipelines consisting of combinations of three read aligners (BWA-MEM, Bowtie2, Novoalign) and four variant callers (GATK HaplotypeCaller, SAMtools mpileup, Freebayes, Ion Proton Variant Caller) against a dataset of highly confident "gold standard" human variants published by the 1000 Genomes Project. In that study, the combination of BWA-MEM with SAMtools achieved the greatest accuracy in SNP identification. The two pipelines using these tools in our study (IGST and GB-eaSy) attained the greatest allelic concordance with WGS in the six lines studied.

Each of the other three pipelines investigated here uses a different variant caller. TASSEL-GBS, which calls SNPs using its own binomial likelihood ratio method (Glaubitz et al. 2014), also agreed well with WGS SNP calls. However, because it found fewer SNPs overall, TASSEL-GBS' number of validated SNPs was lower than that of GB-eaSy and IGST. Stacks uses a multinomial-based likelihood model for SNP calling, which produced an allelic agreement above 95% but the fewest validated SNPs in each line due in part to its finding fewer SNPs overall. Stacks' variant caller consults the reference genome only for read placement, not for nucleotide comparisons, as it is optimized for high-coverage analysis of biological diversity RAD sequencing experiments in which reference genomes are often not available (Catchen et al. 2013). For the low-coverage data typical of plant breeding studies, it is likely a disadvantage that Stacks does not utilize the Bayesian priors available from high-quality reference genomes. However, for organisms lacking a reference genome, the Stacks approach is likely optimal. Finally, Fast-GBS' variant caller, Platypus, uses a haplotype-based strategy to identify variants. A previous analysis

(Torkamaneh et al. 2016) found that comparison of Fast-GBS SNP calls with WGS data in soybean yielded an accuracy of 98.7%, a result consistent with those presented here. Platypus' superiority in indel identification but comparatively lower performance in SNP calling has been reported (Tian et al. 2016), which may explain its slightly lower agreement with WGS compared to the tools used in TASSEL-GBS, IGST and GB-eaSy.

Across all six soybean lines examined, GB-eaSy, TASSEL-GBS and IGST identified SNPs with the greatest accuracy (over 99%), based on comparison to WGS SNPs called by GATK HaplotypeCaller. The accuracy of Fast-GBS and Stacks was lower but still reasonably high (never below 97%). This high accuracy among all five workflows, coupled with the low SNP convergence between them, indicates that they arrived at largely complementary sets of valid SNP calls. For instance, GB-eaSy, TASSEL-GBS and IGST converged on just 2501 (12.85%) of their total 19465 unique SNPs found in Prize. Similarly, these three pipelines converged on just 6781 (17.02%) of their 39853 unique SNPs found in Skylla. These results echo a previous report on barley GBS data in which approximately half of SNPs called by TASSEL-GBS and BCFtools/SAMtools were unique to each pipeline (Mascher et al. 2013).

Storage, run time and ease of use

TASSEL-GBS, the workflow with the smallest storage requirements, used approximately half of the hard disk space required by Stacks and GB-eaSy. While it used the least disk space, TASSEL-GBS identified the fewest SNPs. Both IGST and Fast-GBS found more SNPs than TASSEL-GBS but required the largest amount of disk space due to their generation of many uncompressed intermediate files, even with parameters set to delete intermediate files where possible. This characteristic could hinder their adoption by users with limited computer storage capacity. Across pipelines, these patterns also emerged in run time differences, which may be determined to a large extent by read-write rather than CPU operations. IGST and Stacks required considerably more time to run than TASSEL-GBS, Fast-GBS and GB-eaSy. For instance, IGST needed over 18 h to process data from Population 2, while TASSEL-GBS finished in less than 5 h. Long completion times limit the throughput of data processing, making the slower pipelines less suitable for time-sensitive projects. GB-eaSy's run times were intermediate, ranking ahead of IGST, Stacks and occasionally Fast-GBS but behind TASSEL-GBS.

Given the complexities of GBS analysis, a critical element of any bioinformatics pipeline is ease of use. The five analysis platforms in this study rely on two command input strategies. In TASSEL-GBS and Stacks, the user inputs individual commands that each run a different step of the pipeline. In contrast, IGST, Fast-GBS and GB-eaSy automate this process by requiring just one command from the user to execute all steps; however, IGST and Fast-GBS also depend on adherence to a rigid convention for file naming and directory structure to ensure successful completion. GB-eaSy does not require the user to follow strict instructions for setting up directory structure or naming files. Instead, it uses a parameters file to customize the analysis for each project based on user input.

Another consideration in ease of use is the ability of a method to carry out all the steps necessary to produce accurate SNP calls. For our data, TASSEL-GBS and Fast-GBS required extra steps not built into their pipelines to improve the accuracy of their SNP calls. Fast-GBS initially appeared to identify significantly fewer SNPs than the other methods and showed lower agreement with WGS. However, after decomposition of compound SNPs into allelic primitives using the VariantsToAllelicPrimitives script in GATK, the apparent performance of Fast-GBS improved considerably; these optimized results were used in the comparisons. Prior to running TASSEL, we removed adapter sequence from the reads using Cutadapt, adding an additional step to the workflow, after finding that adapter contamination significantly impaired the accuracy of TASSEL-GBS SNP calls. Again, the optimized results after the trimming step were used in the comparisons. In GB-eaSy, these additional steps either are not required or are built into the pipeline itself.

CONCLUSIONS

Here we introduced the GB-eaSy pipeline and compared its performance to four other GBS workflows and to whole-genome sequencing on low-coverage data from soybean. Differences were apparent between the performance of these methods depending on the aims of the developers. TASSEL-GBS was designed for plant breeding applications and to run on individual PCs, and is thus optimized for maximum computational efficiency. The compromises inherent in the tag strategy limit the number of SNPs that TASSEL-GBS can identify using

datasets such as those utilized here. Stacks is a method developed primarily for high-depth RAD sequencing on organisms without reference genomes. It is likely to be an excellent choice for breeders in orphan crops, as well as for biological diversity applications, but the reference-genome independence of the variant calling algorithm and the low-coverage data used here render the current version less accurate than methods incorporating reference sequences for low-depth GBS in soybean. Fast-GBS and IGST are, like GB-easy, methods designed for plant breeding applications on complex crops with high-quality reference genomes. The overall performance of these methods in terms of SNP number and accuracy is similar. GB-easy has an advantage over the other methods in terms of resources needed (particularly disk space), ease of implementation, and number of accurate SNPs identified. Although our results demonstrate relatively low SNP concordance between GBS pipelines, comparison of each GBS pipeline to WGS data indicates that the SNP calls from each are highly accurate, particularly those generated by GB-eaSy, TASSEL-GBS and IGST. These findings suggest that a comprehensive approach integrating the results from multiple GBS analysis methods may be the optimal strategy to obtain the largest, most highly accurate SNP yield possible from low-coverage polyploid sequence data.

TABLES AND FIGURES

	Population 1	Population 2	Population 3
Description	F2 from cross between Prize and mutagenized Williams 82	F2 from cross between two breeding lines	81 unrelated lines
Number of samples	378	391	200
Sequencer	Illumina HiSeq2500	Illumina HiSeq4000	Illumina HiSeq2500
Read length	100 bp	100 bp	100 bp
Number of reads	234,574,472 (single-end)	392,001,642 (single-end)	247,063,538 (single-end)
Average depth per sequenced base	1.87 reads	3.63 reads	4.47 reads
Average percent of genome covered by at least 1 read	2.29	2.02	2.35
Average percent of genome covered by at least 2 reads	1.08	1.42	1.71

Table 3.1. GBS library data for the three populations analyzed in this study.

	TASSEL-GBS	IGST	Fast-GBS	Stacks	GB-eaSy
Demultiplex reads	GBSSeqToTagDBPlugin, TagExportToTagDBPlugin	Sabre	Sabre	process_radtags	GBSX
Trim adapters	cutadapt*	trimAdaptor3.py	cutadapt	process_radtags	GBSX
Align to reference	bwa-mem*	bwa-aln	bwa-mem	bwa-mem*	bwa-mem
Call SNPs	DiscoverySNPCallerPluginV2, ProductionSNPCallerPluginV2	SAMtools/ BCFtools	Platypus	pstacks, cstacks, stacks, populations	BCFtools

Table 3.2. Major steps of the 5 GBS workflows analyzed. Each workflow uses a different series of tools to carry out read demultiplexing, adapter trimming, alignment to the reference genome, and SNP calling. Asterisks indicate steps performed manually outside of the workflow.

	Prize	LG12	Magellan	Maverick	Prohio	Skylla
Population of origin	Population 1	Population 2	Population 3	Population 3	Population 3	Population 3
Read length	100 bp	150 bp	150 bp	150 bp	150 bp	150 bp
Number of reads	130,404,160 (paired-end)	43,756,742 (paired-end)	12,880,066 (paired-end)	19,038,600 (paired-end)	34,177,159 (paired-end)	23,190,927 (paired-end)
Coverage (LN / G)	13.65	6.87	2.02	2.99	5.37	3.64
Percent of genome covered by at least 1 read	98.67	97.76	74.38	94.06	98.36	96.16
Percent of genome covered by at least 2 reads	98.31	97.04	73.03	85.18	97.27	90.36

Table 3.3. WGS library data for six lines. Prize and LG12 were also included in GBS Populations 1 and 2, respectively. Magellan, Maverick, Prohio and Skylla were included in GBS Population 3. Coverage was computed as the product of read length and number of reads, divided by genome size.

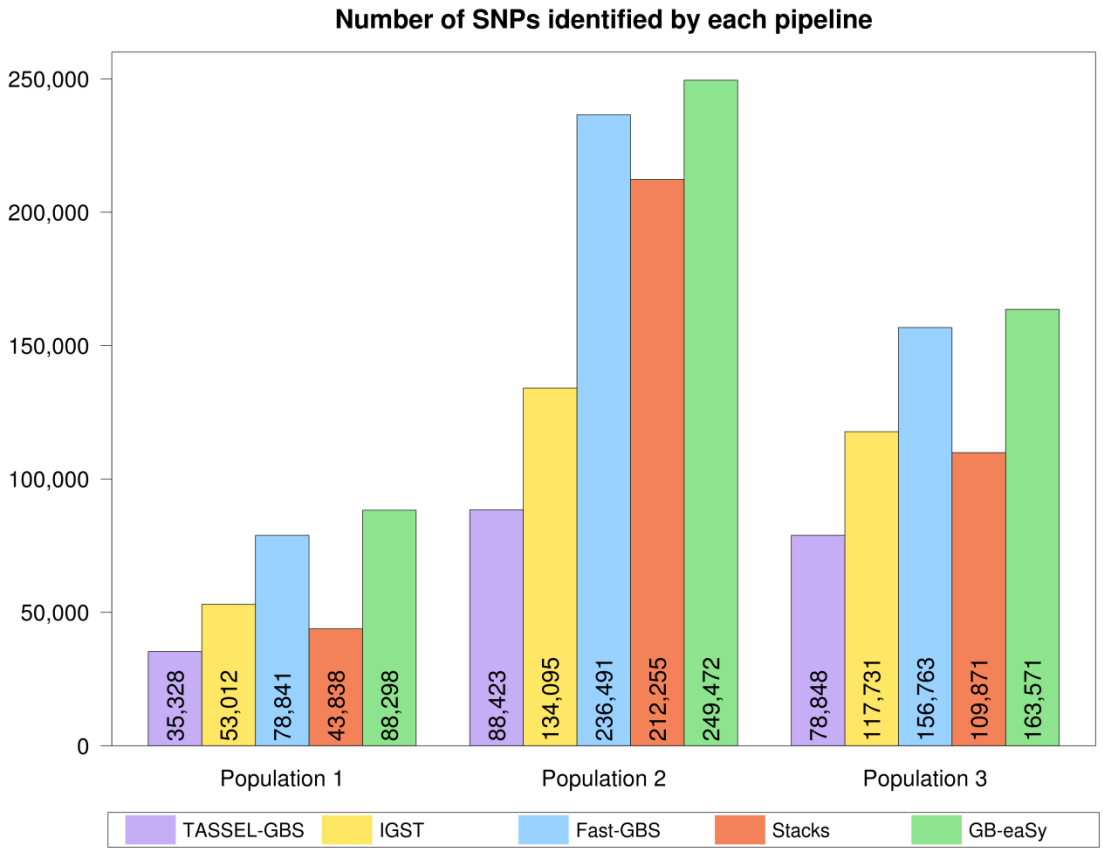

Figure 3.1. Number of SNPs identified by each pipeline in 3 populations. SNPs with a minimum read depth of 2 reads are shown.

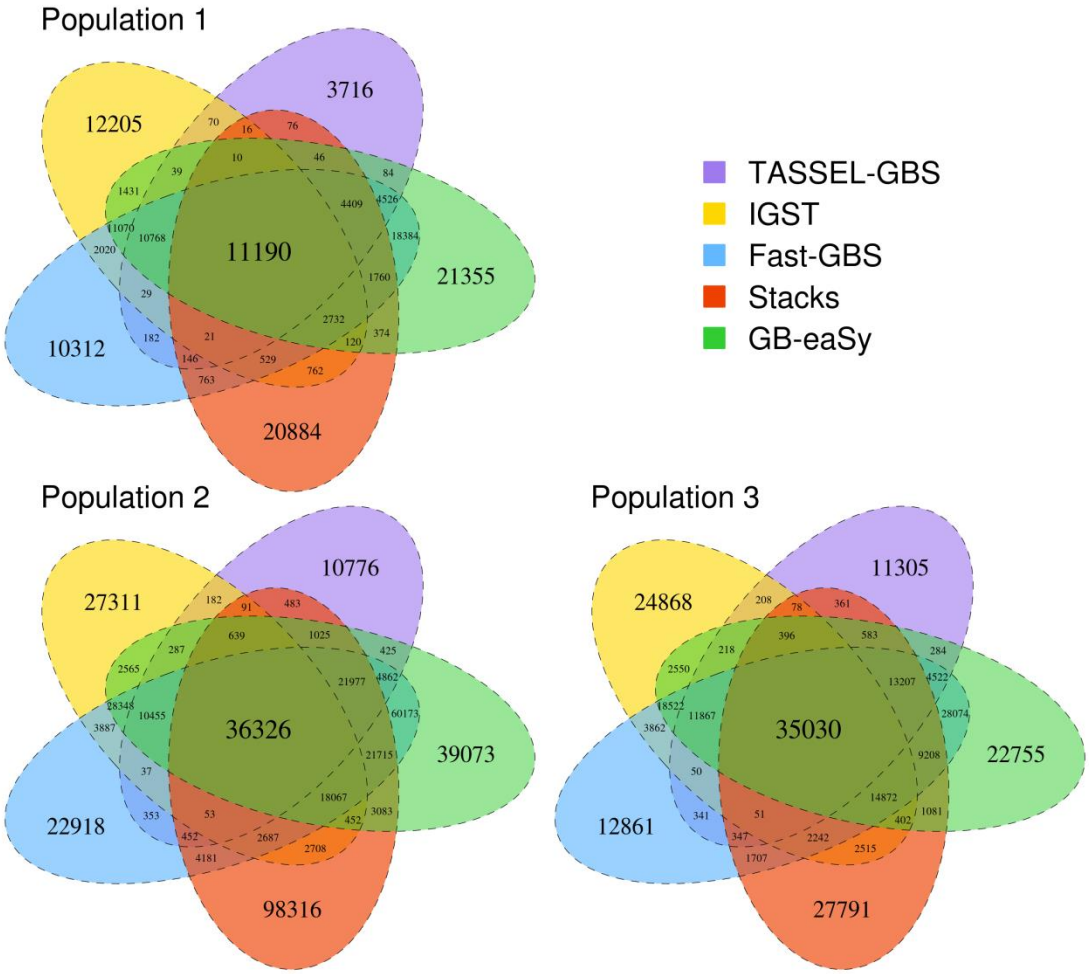

Figure 3.2A. SNP overlap among 5 GBS pipelines. SNPs with a minimum read depth of 2 reads are shown. All SNPs were called relative to the Williams 82 reference genome.

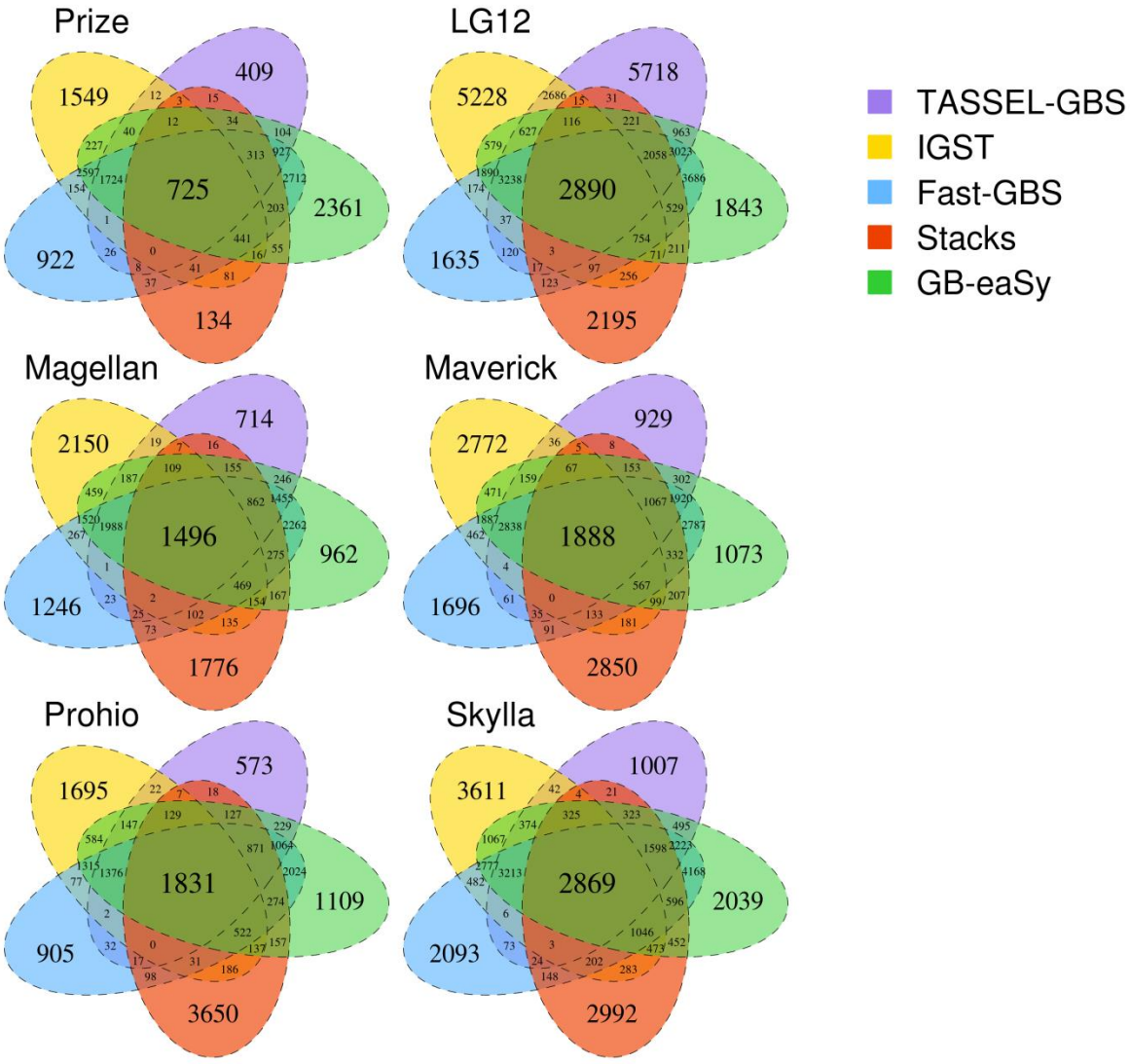

Figure 3.2B. SNP overlap among 5 GBS pipelines for 6 lines from 3 populations. Prize is from GBS Population 1, LG12 is from GBS Population 2, and the four remaining lines are from GBS Population 3. SNPs with a minimum depth of 2 reads are shown. All SNPs were called relative to the Williams 82 reference genome.

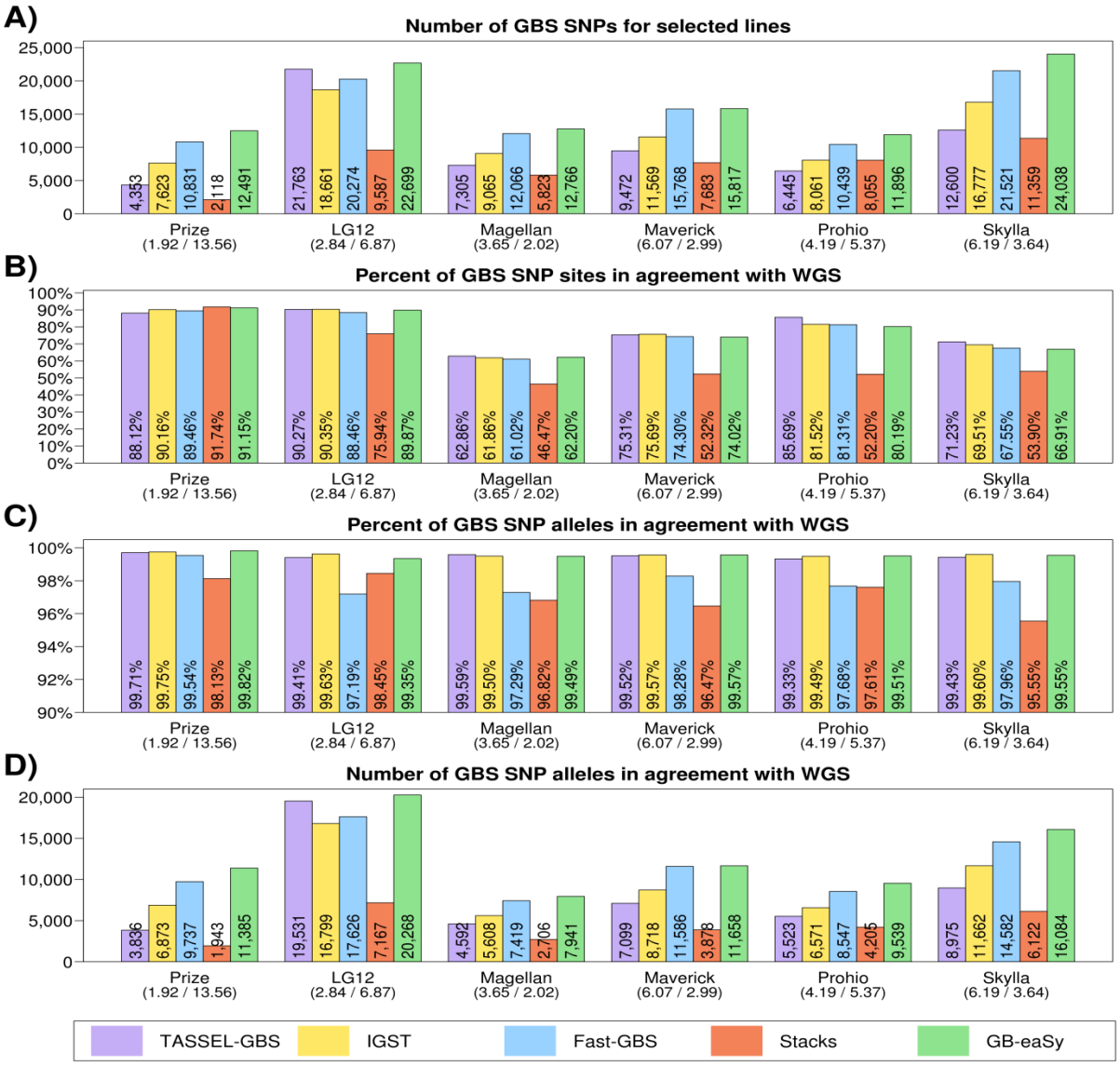

Figure 3.3. Comparisons between GBS SNPs and WGS SNPs for 6 individual soybean lines. Prize is from GBS Population 1, LG12 is from GBS Population 2, and the four remaining lines are from GBS Population 3. Panel A shows the total number of SNPs identified in each line by 5 GBS pipelines. Panel B shows the percent of GBS SNP sites from panel A in agreement with WGS for each line. Panel C and D show the percent and number (respectively) of GBS SNP alleles from panel A in agreement with WGS. SNPs with a minimum read depth of 2 reads are shown. Below each soybean line is shown its average depth of sequenced GBS bases followed by its WGS coverage. All SNPs were called relative to the Williams 82 reference genome.

Population 1:

	TASSEL	IGST	Fast-GBS	Stacks	GB-eaSy
Missing data per line	84.5%	85.4%	85.0%	89.7%	83.4%
SNPs in 25% of lines	6812	12334	18731	3576	23633
SNPs in 50% of lines	1237	1714	2984	202	3558
SNPs in 75% of lines	736	112	382	31	407
SNPs in 90% of lines	335	25	75	2	119

Population 2:

	TASSEL	IGST	Fast-GBS	Stacks	GB-eaSy
Missing data per line	59.4%	70.8%	70.0%	66.1%	71.5%
SNPs in 25% of lines	65119	68805	122801	142154	120437
SNPs in 50% of lines	35107	39055	76485	52991	76717
SNPs in 75% of lines	2185	1548	4418	372	4880
SNPs in 90% of lines	973	26	219	21	187

Population 3:

	TASSEL	IGST	Fast-GBS	Stacks	GB-eaSy
Missing data per line	62.4%	69.3%	68.4%	67.2%	69.6%
SNPs in 25% of lines	54960	65695	88904	69300	88025
SNPs in 50% of lines	18859	22369	32077	19756	32698
SNPs in 75% of lines	6196	7813	12204	4539	13005
SNPs in 90% of lines	775	479	934	98	1352

Table 3.4. Missing data fraction generated by each GBS pipeline. The average percent of missing data per line is shown, as well as the number of SNPs detected at various proportions within each population.

	TASSEL	IGST	Fast-GBS	Stacks	GB-eaSy
Population 1	2:08	12:17	3:20	8:36	5:21
Population 2	4:58	18:46	8:01	16:34	6:51
Population 3	3:38	11:28	4:06	10:15	4:23

Table 3.5. Wall-clock time to completion for each GBS pipeline (h:mm).

	TASSEL	IGST	Fast-GBS	Stacks	GB-eaSy
Population 1	22.4	347.6	190	43.6	50.6
Population 2	47.6	262.6	379	87.6	78.6
Population 3	41.6	165.6	264	69.6	65.6

Table 3.6. Disk space required for each GBS pipeline (GB).

REFERENCES

Alzheimer's Association. 2018. "2018 Alzheimer's Disease Facts and Figures." *Alzheimers Dement* 14 (3): 367–429. https://doi.org/10.1016/j.jalz.2018.02.001.

Andersen, Olav M. Ina-Maria Rudolph, Thomas E. Willnow. 2016 "Risk Factor SORL1: From Genetic Association to Functional Validation in Alzheimer's Disease." Acta Neuropathol 132: 653-665

Baes, Christine F, Marlies A Dolezal, James E Koltes, Beat Bapst, Eric Fritz-Waters, Sandra Jansen, Christine Flury, et al. 2014. "Evaluation of Variant Identification Methods for Whole Genome Sequencing Data in Dairy Cattle." *BMC Genomics* 15 (1): 948. https://doi.org/10.1186/1471-2164-15-948.

Baird, Nathan A., Paul D. Etter, Tressa S. Atwood, Mark C. Currey, Anthony L. Shiver, Zachary A. Lewis, Eric U. Selker, William A. Cresko, and Eric A. Johnson. 2008. "Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers." *PLoS ONE* 3 (10): e3376. https://doi.org/10.1371/journal.pone.0003376.

Beecham, Gary W., J.C. Bis, E.R. Martin, S.-H. Choi, A. L. DeStefano, C.M. van Duijn, M. Fornage, et al. 2017. "The Alzheimer's Disease Sequencing Project: Study Design and Sample Selection." *Neurology Genetics* 3 (5): e194. https://doi.org/10.1212/NXG.0000000000000194.

Bis, Joshua C., Alzheimer's Disease Sequencing Project, Xueqiu Jian, Brian W. Kunkle, Yuning Chen, Kara L. Hamilton-Nelson, William S. Bush, et al. 2018. "Whole Exome Sequencing Study Identifies Novel Rare and Common Alzheimer's-Associated Variants Involved in Immune Response and Transcriptional Regulation." *Molecular Psychiatry*, August. https://doi.org/10.1038/s41380-018-0112-7.

Blanc, G. 2004. "Widespread Paleopolyploidy in Model Plant Species Inferred from Age Distributions of Duplicate Genes." *The Plant Cell* 16 (7): 1667–78. https://doi.org/10.1105/tpc.021345.

Bolger, Marie, Rainer Schwacke, Heidrun Gundlach, Thomas Schmutzer, Jinbo Chen, Daniel Arend, Markus Oppermann, et al. 2017. "From Plant Genomes to Phenotypes." *Journal of Biotechnology* 261 (November): 46–52. https://doi.org/10.1016/j.jbiotec.2017.06.003.

Braak, H., and E. Braak. 1991. "Neuropathological Stageing of Alzheimer-Related Changes." *Acta Neuropathologica* 82 (4): 239–59. https://doi.org/10.1007/BF00308809.

Bradbury, P. J., Z. Zhang, D. E. Kroon, T. M. Casstevens, Y. Ramdoss, and E. S. Buckler. 2007. "TASSEL: Software for Association Mapping of Complex Traits in Diverse Samples." *Bioinformatics* 23 (19): 2633–35. https://doi.org/10.1093/bioinformatics/btm308.

Catchen, Julian, Paul A. Hohenlohe, Susan Bassham, Angel Amores, and William A. Cresko. 2013. “Stacks: An Analysis Tool Set for Population Genomics.” *Molecular Ecology* 22 (11): 3124–40. https://doi.org/10.1111/mec.12354.

Chamary, J. V., Joanna L. Parmley, and Laurence D. Hurst. 2006. “Hearing Silence: Non-Neutral Evolution at Synonymous Sites in Mammals.” *Nature Reviews Genetics* 7 (2): 98–108. https://doi.org/10.1038/nrg1770.

Chang, Christopher C, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. 2015. “Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets.” *GigaScience* 4 (1). https://doi.org/10.1186/s13742-015-0047-8.

Cingolani, Pablo, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J. Land, Xiangyi Lu, and Douglas M. Ruden. 2012. “A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff: SNPs in the Genome of Drosophila Melanogaster Strain w ^{1118} ; Iso-2; Iso-3.” *Fly* 6 (2): 80–92. https://doi.org/10.4161/fly.19695.

Claros, Manuel Gonzalo, Rocío Bautista, Darío Guerrero-Fernández, Hicham Benzerki, Pedro Seoane, and Noé Fernández-Pozo. 2012. “Why Assembling Plant Genome Sequences Is So Challenging.” *Biology* 1 (2): 439–59. https://doi.org/10.3390/biology1020439.

Clevenger, Josh P., and Peggy Ozias-Akins. 2015. “SWEEP: A Tool for Filtering High-Quality SNPs in Polyploid Crops.” *G3: Genes|Genomes|Genetics* 5 (9): 1797–1803. https://doi.org/10.1534/g3.115.019703.

Corder, E., A. Saunders, W. Strittmatter, D. Schmechel, P. Gaskell, G. Small, A. Roses, J. Haines, and M. Pericak-Vance. 1993. “Gene Dose of Apolipoprotein E Type 4 Allele and the Risk of Alzheimer’s Disease in Late Onset Families.” *Science* 261 (5123): 921–23. https://doi.org/10.1126/science.8346443.

Cuyvers, Elise, and Kristel Sleegers. 2016. “Genetic Variations Underlying Alzheimer’s Disease: Evidence from Genome-Wide Association Studies and Beyond.” *The Lancet Neurology* 15 (8): 857–68. https://doi.org/10.1016/S1474-4422(16)00127-7.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, et al. 2011. “The Variant Call Format and VCFtools.” *Bioinformatics* 27 (15): 2156–58. https://doi.org/10.1093/bioinformatics/btr330.

Del Prete, Dolores, Jan M. Suski, Bénédicte Oulès, Delphine Debayle, Anne Sophie Gay, Sandra Lacas-Gervais, Renaud Bussiere, et al. 2016. “Localization and Processing of the Amyloid-β Protein Precursor in Mitochondria-Associated Membranes.” *Journal of Alzheimer’s Disease* 55 (4): 1549–70. https://doi.org/10.3233/JAD-160953.

Drummond, D. Allan, and Claus O. Wilke. 2008. “Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution.” *Cell* 134 (2): 341–52. https://doi.org/10.1016/j.cell.2008.05.042.

Elshire, Robert J., Jeffrey C. Glaubitz, Qi Sun, Jesse A. Poland, Ken Kawamoto, Edward S. Buckler, and Sharon E. Mitchell. 2011. “A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species.” *PLoS ONE* 6 (5): e19379. https://doi.org/10.1371/journal.pone.0019379.

Ertekin-Taner, Nilüfer. 2007. “Genetics of Alzheimer’s Disease: A Centennial Review.” *Neurologic Clinics* 25 (3): doi:10.1016/j.ncl.2007.03.009. Genetics. https://doi.org/10.1016/j.asieco.2008.09.006.EAST.

Fan, Yanjie, Wu Yin, Bing Hu, Antonie D. Kline, Victor Wei Zhang, Desheng Liang, Yu Sun, et al. 2018. “De Novo Mutations of CCNK Cause a Syndromic Neurodevelopmental Disorder with Distinctive Facial Dysmorphism.” *The American Journal of Human Genetics* 103 (3): 448–55. https://doi.org/10.1016/j.ajhg.2018.07.019.

Farrer, Lindsay A, L Adrienne Cupples, Jonathan L Haines, Bradley Hyman, Walter A Kukull, Richard Mayeux, Richard H Myers, Margaret A Pericak-Vance, Neil Risch, and Cornelia M van Duijn. 1997. “Effects of Age , Sex , and Ethnicity on the Association Between Apolipoprotein E Genotype and Alzheimer Disease.” *Jama* 278: 1349–56.

Feschotte, Cédric, Ning Jiang, and Susan R. Wessler. 2002. “Plant Transposable Elements: Where Genetics Meets Genomics.” *Nature Reviews Genetics* 3 (5): 329–41. https://doi.org/10.1038/nrg793.

Frisoni, Giovanni B., Marina Boccardi, Frederik Barkhof, Kaj Blennow, Stefano Cappa, Konstantinos Chiotis, Jean Francois Démonet, et al. 2017. “Strategic Roadmap for an Early Diagnosis of Alzheimer’s Disease Based on Biomarkers.” *The Lancet Neurology* 16 (8): 661–76. https://doi.org/10.1016/S1474-4422(17)30159-X.

Furuta, Tomoyuki, Motoyuki Ashikari, Kshirod K Jena, Kazuyuki Doi, and Stefan Reuscher. 2017. “Adapting Genotyping-by-Sequencing for Rice F2 Populations.” G3 Genes|Genomes|Genetics. 2017;7. doi:10.1534/g3.116.038190.

Gatz, Margaret, Chandra a Reynolds, Laura Fratiglioni, Boo Johansson, James a Mortimer, Stig Berg, Amy Fiske, and Nancy L Pedersen. 2006. “Role of Genes and Environments for Explaining Alzheimer’s Disease.” *Archives of General Psychiatry* 63 (2): 168–74. https://doi.org/10.1001/archpsyc.63.2.168.

Glaubitz, Jeffrey C., Terry M. Casstevens, Fei Lu, James Harriman, Robert J. Elshire, Qi Sun, and Edward S. Buckler. 2014. “TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline.” *PLoS ONE* 9 (2): e90346. https://doi.org/10.1371/journal.pone.0090346.

Griseri, Paola, Christine Bourcier, Corinne Hieblot, Khadija Essafi-Benkhadir, Emmanuel Chamorey, Christian Touriol, and Gilles Pagès. 2011. “A Synonymous Polymorphism of the Tristetraprolin (TTP) Gene, an AU-Rich MRNA-Binding Protein, Affects Translation Efficiency and Response to Herceptin Treatment in Breast Cancer Patients.” *Human Molecular Genetics* 20 (23): 4556–68. https://doi.org/10.1093/hmg/ddr390.

Gu, Wanjun, Tong Zhou, and Claus O. Wilke. 2010. “A Universal Trend of Reduced MRNA Stability near the Translation-Initiation Site in Prokaryotes and Eukaryotes.” *PLoS Computational Biology* 6 (2): e1000664. doi:10.1371/journal.pcbi.1000664. https://doi.org/10.1371/journal.pcbi.1000664.

Guo, Tong, Wendy Noble, and Diane P. Hanger. 2017. “Roles of Tau Protein in Health and Disease.” *Acta Neuropathologica* 133 (5): 665–704. https://doi.org/10.1007/s00401-017-1707-9.

Hebert, Liesi E., Jennifer Weuve, Paul A Scherr, and Denis A Evans. 2013. “Alzheimer Disease in the United States (2010 – 2050) Estimated Using the 2010 Census.” *Neurology* 80: 1778–83. https://doi.org/10.1212/WNL.0b013e31828726f5.

Herten, Koen, Matthew S Hestand, Joris R Vermeesch, and Jeroen KJ Van Houdt. 2015. “GBSX: A Toolkit for Experimental Design and Demultiplexing Genotyping by Sequencing Experiments.” *BMC Bioinformatics* 16 (1): 73. https://doi.org/10.1186/s12859-015-0514-3.

Holstege, Henne, Sven J van der lee, Marc Hulsman, Tsz Hang Wong, Jeroen GJ van Rooij, Marjan Weiss, Eva Louwersheimer et al. 2017. “Characterization of Pathogenic SORL1 Genetic Variants For Association with Alzheimer’s Disease: A Clinical Interpretation Atrategy.” European Journal of Human Genetics 25: 973-981.

Hunt, Ryan C., Vijaya L. Simhadri, Matthew Iandoli, Zuben E. Sauna, and Chava Kimchi-Sarfaty. 2014a. “Exposing Synonymous Mutations.” *Trends in Genetics* 30 (7): 308–21. https://doi.org/10.1016/j.tig.2014.04.006.

Hwang, Sohyun, Eiru Kim, Insuk Lee, and Edward M. Marcotte. 2016. “Systematic Comparison of Variant Calling Pipelines Using Gold Standard Personal Exome Variants.” *Scientific Reports* 5 (1): 17875. https://doi.org/10.1038/srep17875.

Hyman, Bradley T., Creighton H. Phelps, Thomas G. Beach, Eileen H. Bigio, Nigel J. Cairns, Maria C. Carrillo, Dennis W. Dickson, et al. 2013. “National Insitute on Aging- Alzheimer’s Association Guidelines for the Neuropathologic Assessment of Alzheimer’s Disease.” *Alzheimer’s & Dementia* 8 (1): 1–13. https://doi.org/10.1016/j.jalz.2011.10.007.National.

Ikemura, Toshimichi. 1985. “Codon Usage and TRNA Content in Unicellular and Multicellular Organisms.” *Molecular Biology and Evolution* 2 (June): 13–34. https://doi.org/10.1093/oxfordjournals.molbev.a040335.

Im, Eu Hyun, Yoonsoo Hahn, and Sun Shim Choi. 2018. “Functional Relevance of Synonymous Alleles Reflected in Allele Rareness in the Population.” *Genomics*, doi:10.1016/j.ygeno.2018.04.003. https://doi.org/10.1016/j.ygeno.2018.04.003.

Iqbal, Khalid, Fei Liu, Cheng-Xin Gong, and Inge Grundke-Iqbal. 2010. “Tau in Alzheimer Disease and Related Tauopathies.” *Current Alzheimer Research* 6 (8): 656–64. https://doi.org/10.1111/j.1743-6109.2008.01122.x.Endothelial.

Jack, Clifford R, David S Knopman, William J Jagust, Leslie M Shaw, Paul S Aisen, W Weiner, Ronald C Petersen, and John Q Trojanowski. 2010. “Hypothetical Model of Dynamic Biomarkers of Alzheimer’s Pathological Cascade.” *Lancet Neurology* 9 (1): doi:10.1016/S1474-4422(09)70299-6. https://doi.org/10.1016/S1474-4422(09)70299-6.Hypothetical.

Jacobo, S. M. P., M. M. DeAngelis, I. K. Kim, and A. Kazlauskas. 2013. “Age-Related Macular Degeneration-Associated Silent Polymorphisms in HtrA1 Impair Its Ability To Antagonize Insulin-Like Growth Factor 1.” *Molecular and Cellular Biology* 33 (10): 1976–90. https://doi.org/10.1128/MCB.01283-12.

Jiao, Wen-Biao, and Korbinian Schneeberger. 2017. “The Impact of Third Generation Genomic Technologies on Plant Genome Assembly.” *Current Opinion in Plant Biology* 36 (April): 64–70. https://doi.org/10.1016/j.pbi.2017.02.002.

Jun, G., Matthew Flickinger, Kurt N. Hetrick, Jane M. Romm, Kimberly F. Doheny, Goncalo R. Abecasis, Michael Boehnke & Hyun Min Kang. 2012. “Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-based Genotype Data.” *American Journal of Human Genetics* 91: 839-848.

Kim, Kyung Do, Jin Hee Shin, Kyujung Van, Dong Hyun Kim, Suk-Ha Lee. 2009. “Dynamic Rearrangements Determine Genome Organization and Useful Traits in Soybean.” *Plant Physiology* 151:1066-1076.

Kirchner, Sebastian, Zhiwei Cai, Robert Rauscher, Nicolai Kastelic, Melanie Anding, Andreas Czech, Bertrand Kleizen, et al. 2017. “Alteration of Protein Function by a Silent Polymorphism Linked to TRNA Abundance.” *PLoS Biology* 15 (5): e2000779, doi:10.1371/journal.pbio.2000779. https://doi.org/10.1371/journal.pbio.2000779.

Kyriakidou, Maria, Helen H. Tai, Noelle L. Anglin, David Ellis, and Martina V. Strömvik. 2018. “Current Strategies of Polyploid Plant Genome Sequence Assembly.” *Frontiers in Plant Science* 9 (November): 1660. https://doi.org/10.3389/fpls.2018.01660.

Lambert, Jean-Charles. 2013. “Meta-Analysis of 74,046 Individuals Identifies 11 New Susceptibility Loci for Alzheimer’s Disease.” *Nature Genetics* 45 (12): 1452–58. https://doi.org/10.1038/ng.2802.Meta-analysis.

Langmead, Ben, Cole Trapnell, Mihai Pop, and Steven L Salzberg. 2009. “Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome.” *Genome Biology* 10 (3): R25. https://doi.org/10.1186/gb-2009-10-3-r25.

Lavner, Yizhar, and Daniel Kotlar. 2005. “Codon Bias as a Factor in Regulating Expression via Translation Rate in the Human Genome.” *Gene* 345 (1 SPEC. ISS.): 127–38. https://doi.org/10.1016/j.gene.2004.11.035.

Li, H. 2011. “A Statistical Framework for SNP Calling, Mutation Discovery, Association Mapping and Population Genetical Parameter Estimation from Sequencing Data.” *Bioinformatics* 27 (21): 2987–93. https://doi.org/10.1093/bioinformatics/btr509.

Li, H., and R. Durbin. 2009. “Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform.” *Bioinformatics* 25 (14): 1754–60. https://doi.org/10.1093/bioinformatics/btp324.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics* 25 (16): 2078–79. https://doi.org/10.1093/bioinformatics/btp352.

Li, Heng. 2013. “Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM.” *ArXiv:1303.3997 [q-Bio]*, March. http://arxiv.org/abs/1303.3997.

Li, Mingyao, Muredach P. Reilly, Daniel J. Rader, and Li-San Wang. 2010. “Correcting Population Stratification in Genetic Association Studies Using a Phylogenetic Approach.” *Bioinformatics* 26 (6): 798–806. https://doi.org/10.1093/bioinformatics/btq025.

Liao, Jung-Chi, T. Tony Yang, Rueyhung Roc Weng, Ching-Te Kuo, and Chih-Wei Chang. 2015. “TTBK2: A Tau Protein Kinase beyond Tau Phosphorylation.” *BioMed Research International* 2015: 1–10. https://doi.org/10.1155/2015/575170.

Liu, Chia-Chen, Takahisa Kanekiyo, Huaxi Xu, and Guojun Bu. 2013. “Apolipoprotein E and Alzheimer Disease: Risk, Mechanisms, and Therapy.” *Nature Reviews Neurology* 9 (2): 106–18. https://doi.org/10.1016/j.asieco.2008.09.006.EAST.

Liu, Hui, Micha Bayer, Arnis Druka, Joanne R Russell, Christine A Hackett, Jesse Poland, Luke Ramsay, Pete E Hedley, and Robbie Waugh. 2014. “An Evaluation of Genotyping by Sequencing (GBS) to Map the Breviaristatum-e (Ari-e) Locus in Cultivated Barley.” *BMC Genomics* 15 (1): 104. https://doi.org/10.1186/1471-2164-15-104.

Liu, Y, Ke-Zhen Yang, Xiao-Xin Wei, and Xiao-Quan Wang. 2016. "Revisiting the Phosphatidylethanolamine-Binding Protein (PEBP) Gene Family Reveals Cryptic *FLOWERING LOCUS T* Gene Homologs in Gymnosperms and Sheds New Light on Functional Evolution." *New Phytologist* 212 (3): 730–44. https://doi.org/10.1111/nph.14066.

Lord, Jenny, Alexander J. Lu, and Carlos Cruchaga. 2014. "Identification of Rare Variants in Alzheimer's Disease." *Frontiers in Genetics* 5 (OCT): 1–9. https://doi.org/10.3389/fgene.2014.00369.

Lu, Fei, Maria C. Romay, Jeffrey C. Glaubitz, Peter J. Bradbury, Robert J. Elshire, Tianyu Wang, Yu Li, et al. 2015. "High-Resolution Genetic Mapping of Maize Pan-Genome Sequence Anchors." *Nature Communications* 6 (1): 6914. https://doi.org/10.1038/ncomms7914.

Marouli, Eirini, Mariaelisa Graff, Carolina Medina-Gomez, Ken Sin Lo, Andrew R. Wood, Troels R. Kjaer, Rebecca S. Fine, et al. 2017. "Rare and Low-Frequency Coding Variants Alter Human Adult Height." *Nature* 542 (7640): 186–90. https://doi.org/10.1038/nature21039.

Martin, Eden R., Ilker Tunc, Zhi Liu, Susan H. Slifer, Ashley H. Beecham, and Gary W. Beecham. 2018. "Properties of Global- and Local-Ancestry Adjustments in Genetic Association Tests in Admixed Populations." *Genetic Epidemiology* 42 (2): 214–29. https://doi.org/10.1002/gepi.22103.

Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;17. doi:10.14806/ej.17.1.200.

Mascher, Martin, Shuangye Wu, Paul St. Amand, Nils Stein, and Jesse Poland. 2013. "Application of Genotyping-by-Sequencing on Semiconductor Sequencing Platforms: A Comparison of Genetic and Reference-Based Marker Ordering in Barley." *PLoS ONE* 8 (10): e76925. https://doi.org/10.1371/journal.pone.0076925.

McKhann, Guy M., David S. Knopman, Howard Chertkow, Bradley T. Hyman, Clifford R. Jack, Claudia H. Kawas, William E. Klunk, et al. 2011. "The Diagnosis of Dementia Due to Alzheimer's Disease: Recommendations from the National Institute on Aging-Alzheimer's Association Workgroups on Diagnostic Guidelines for Alzheimer's Disease." *Alzheimer's & Dementia* 7 (3): 263–69. https://doi.org/10.1016/j.jalz.2011.03.005.

Miller, Jason E., Manu K. Shivakumar, Shannon L. Risacher, Andrew J. Saykin, Seunggeun Lee, Kwangsik Nho, Dokyoon Kim, and for the Alzheimer's Disease Neuroimaging Initiative (ADNI). 2018. "Codon Bias among Synonymous Rare Variants Is Associated with Alzheimer's Disease Imaging Biomarker." In *Biocomputing 2018*, 365–76. Kohala Coast, Hawaii, USA: WORLD SCIENTIFIC. https://doi.org/10.1142/9789813235533_0034.

Miller, M. R., J. P. Dunham, A. Amores, W. A. Cresko, and E. A. Johnson. 2007. “Rapid and Cost-Effective Polymorphism Identification and Genotyping Using Restriction Site Associated DNA (RAD) Markers.” *Genome Research* 17 (2): 240–48. https://doi.org/10.1101/gr.5681207.

Mostafavi, Sara, Chris Gaiteri, Sarah E. Sullivan, Charles C. White, Shinya Tasaki, Jishu Xu, Mariko Taga, et al. 2018. “A Molecular Network of the Aging Human Brain Provides Insights into the Pathology and Cognitive Decline of Alzheimer’s Disease.” *Nature Neuroscience* 21 (6): 811–19. https://doi.org/10.1038/s41593-018-0154-9.

Nicholas, J., P. Peterlongo, S. Tempel. 2016. “Finding and characterizing repeats in plant genomes.” In Plant Bioinformatics: Methods and Protocols 1374: 293-338. New York: Humana Press.

Nicolas, G., C. Charbonnier, D. Wallon, O. Quenez, C. Bellenguez, B. Grenier-Boley, S. Rousseau et al. 2016. “SORL1 rare variants: a major risk factor for familial early-onset lzheimer’s disease.” Molecular Psychiatry 21: 831-836.

Nielsen, Rasmus, Joshua S. Paul, Anders Albrechtsen, and Yun S. Song. 2011. “Genotype and SNP Calling from Next-Generation Sequencing Data.” *Nature Reviews Genetics* 12 (6): 443–51. https://doi.org/10.1038/nrg2986.

O’Rawe, Jason, Tao Jiang, Guangqing Sun, Yiyang Wu, Wei Wang, Jingchu Hu, Paul Bodily, et al. 2013. “Low Concordance of Multiple Variant-Calling Pipelines: Practical Implications for Exome and Genome Sequencing.” *Genome Medicine* 5 (3): 28. https://doi.org/10.1186/gm432.

Orr, H. Allen. 1990. “‘Why Polyploidy Is Rarer in Animals Than in Plants’ Revisited.” *The American Naturalist* 136 (6): 759–70.

Patel, Devanshi, Jesse Mez, Badri N. Vardarajan, Lyndsay Staley, Jaeyoon Chung, Xiaoling Zhang, John J. Farrell, et al. 2019. “Association of Rare Coding Mutations With Alzheimer Disease and Other Dementias Among Adults of European Ancestry.” *JAMA Network Open* 2 (3): e191350. https://doi.org/10.1001/jamanetworkopen.2019.1350.

Pearson, Hugh A., and Chris Peers. 2006. “Physiological Roles for Amyloid β Peptides.” *Journal of Physiology* 575 (1): 5–10. https://doi.org/10.1113/jphysiol.2006.111203.

Penke, Botond, Ferenc Bogár, and Lívia Fülöp. 2017. “β-Amyloid and the Pathomechanisms of Alzheimer’s Disease: A Comprehensive View.” *Molecules* 22 (1692): doi:10.3390/molecules22101692. https://doi.org/10.3390/molecules22101692.

Pirooznia, Mehdi, Melissa Kramer, Jennifer Parla, Fernando S Goes, James B Potash, W McCombie, and Peter P Zandi. 2014. “Validation and Assessment of Variant Calling Pipelines for Next-Generation Sequencing.” *Human Genomics* 8 (1): 14. https://doi.org/10.1186/1479-7364-8-14.

Poland, Jesse, Jeffrey Endelman, Julie Dawson, Jessica Rutkoski, Shuangye Wu, Yann Manes, Susanne Dreisigacker, et al. 2012. “Genomic Selection in Wheat Breeding Using Genotyping-by-Sequencing.” *The Plant Genome Journal* 5 (3): 103. https://doi.org/10.3835/plantgenome2012.06.0006.

Ren, Yingxue, Joseph S. Reddy, Cyril Pottier, Vivekananda Sarangi, Shulan Tian, Jason P. Sinnwell, Shannon K. McDonnell, et al. 2018. “Identification of Missing Variants by Combining Multiple Analytic Pipelines.” *BMC Bioinformatics* 19 (1). https://doi.org/10.1186/s12859-018-2151-0.

Rhoads, Anthony, and Kin Fai Au. 2015. “PacBio Sequencing and Its Applications.” *Genomics, Proteomics & Bioinformatics* 13 (5): 278–89. https://doi.org/10.1016/j.gpb.2015.08.002.

Richard, P, G Thomas, M Pascual de Zulueta, J De Gennes, M Thomas, Andre Cassaigne, Gilbert Bereziat, and A Iron. 1994. “Common and Rare Genotypes of Human ApolipoproteinE Determined by Specific Restriction Profiles of Polymerase Chain Reaction-Amplified DNA.” *Molecular Pathology* 40 (1): 24–29.

Richards, Eric, Mark Reichardt, and Sharon Rogers. 2001. “Preparation of Genomic DNA from Plant Tissue.” In *Current Protocols in Molecular Biology*, edited by Frederick M. Ausubel, Roger Brent, Robert E. Kingston, David D. Moore, J.G. Seidman, John A. Smith, and Kevin Struhl, mb0203s27. Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/0471142727.mb0203s27.

Ridge, Perry G., Shubhabrata Mukherjee, Paul K. Crane, and John S.K. Kauwe. 2013. “Alzheimer’s Disease: Analyzing the Missing Heritability.” *PLoS ONE* 8 (11): e79771, doi:10.1371/journal.pone.0079771. https://doi.org/10.1371/journal.pone.0079771.

Rimmer, Andy, Iain Mathieson, Zamin Iqbal, Stephen R F Twigg, Andrew O M Wilkie, Gil McVean, and Gerton Lunter. 2014. “Integrating Mapping-, Assembly- and Haplotype-Based Approaches for Calling Variants in Clinical Sequencing Applications.” *Nature Genetics* 46 (8): 912–18. https://doi.org/10.1038/ng.3036.

Rubinsztein, David C., and Douglas F. Easton. 1999. “Apolipoprotein E Genetic Variation and Alzheimer’s Disease.” *Dementia and Geriatric Cognitive Disorders* 10 (3): 199–209. https://doi.org/10.1159/000017120.

Sauna, Zuben E., and Chava Kimchi-Sarfaty. 2011. “Understanding the Contribution of Synonymous Mutations to Human Disease.” *Nature Reviews Genetics* 12 (10): 683–91. https://doi.org/10.1038/nrg3051.

Scheben, Armin, Jacqueline Batley, and David Edwards. 2017. “Genotyping-by-Sequencing Approaches to Characterize Crop Genomes: Choosing the Right Tool for the Right Application.” *Plant Biotechnology Journal* 15 (2): 149–61. https://doi.org/10.1111/pbi.12645.

Schmutz, Jeremy, Steven B. Cannon, Jessica Schlueter, Jianxin Ma, Therese Mitros, William Nelson, David L. Hyten, et al. 2010. “Genome Sequence of the Palaeopolyploid Soybean.” *Nature* 463 (7278): 178–83. https://doi.org/10.1038/nature08670.

Serrano-Pozo, Alberto, Jing Qian, Sarah E. Monsell, Rebecca A. Betensky, and Bradley T. Hyman. 2015. “APOE E2 Is Associated with Milder Clinical and Pathological Alzheimer Disease: Effects of *APOE* Alleles in AD.” *Annals of Neurology* 77 (6): 917–29. https://doi.org/10.1002/ana.24369.

Sham, Pak C., and Shaun M. Purcell. 2014. “Statistical Power and Significance Testing in Large-Scale Genetic Studies.” *Nature Reviews Genetics* 15 (5): 335–46. https://doi.org/10.1038/nrg3706.

Shi, Yang, Kaoru Yamada, Shane Antony Liddelow, Scott T. Smith, Lingzhi Zhao, Wenjie Luo, Richard M. Tsai, et al. 2017. “ApoE4 Markedly Exacerbates Tau-Mediated Neurodegeneration in a Mouse Model of Tauopathy.” *Nature* 549 (7673): 523–27. https://doi.org/10.1038/nature24016.

Sonah, Humira, Maxime Bastien, Elmer Iquira, Aurélie Tardivel, Gaétan Légaré, Brian Boyle, Éric Normandeau, et al. 2013. “An Improved Genotyping by Sequencing (GBS) Approach Offering Increased Versatility and Efficiency of SNP Discovery and Genotyping.” *PLoS ONE* 8 (1): e54603. https://doi.org/10.1371/journal.pone.0054603.

Sonah, Humira, Louise O’Donoughue, Elroy Cober, Istvan Rajcan, and François Belzile. 2015. “Identification of Loci Governing Eight Agronomic Traits Using a GBS-GWAS Approach and Validation by QTL Mapping in Soya Bean.” *Plant Biotechnology Journal* 13 (2): 211–21. https://doi.org/10.1111/pbi.12249.

Song, Qijian, Long Yan, Charles Quigley, Brandon D. Jordan, Edward Fickus, Steve Schroeder, Bao-Hua Song, et al. 2017. “Genetic Characterization of the Soybean Nested Association Mapping Population.” *The Plant Genome* 10 (2): 0. https://doi.org/10.3835/plantgenome2016.10.0109.

Spencer, Paige S., Efraín Siller, John F. Anderson, and José M. Barral. 2012. “Silent Substitutions Predictably Alter Translation Elongation Rates and Protein Folding Efficiencies.” *Journal of Molecular Biology* 422 (3): 328–35. https://doi.org/10.1016/j.jmb.2012.06.010.

Sperling, Reisa A., Paul S. Aisen, Laurel A. Beckett, David A. Bennett, Suzanne Craft, Anne M. Fagan, Takeshi Iwatsubo, et al. 2011. “Toward Defining the Preclinical Stages of Alzheimer’s Disease: Recommendations from the National Institute on Aging-Alzheimer’s Association Workgroups on Diagnostic Guidelines for Alzheimer’s Disease.” *Alzheimer’s and Dementia* 7 (3): 280–92. https://doi.org/10.1016/j.jalz.2011.03.003.

Tange, Ole. 2011. “GNU Parallel: The Command-Line Power Tool,” 6. https://doi.org/10.5281/zenodo.16303.

Taylor, Laura M., Pamela J. McMillan, Nicole F. Liachko, Timothy J. Strovas, Bernardino Ghetti, Thomas D. Bird, C. Dirk Keene, and Brian C. Kraemer. 2018. “Pathological Phosphorylation of Tau and TDP-43 by TTBK1 and TTBK2 Drives Neurodegeneration.” *Molecular Neurodegeneration* 13 (1). https://doi.org/10.1186/s13024-018-0237-9.

Thankaswamy-Kosalai, Subazini, Partho Sen, and Intawat Nookaew. 2017. “Evaluation and Assessment of Read-Mapping by Multiple next-Generation Sequencing Aligners Based on Genome-Wide Characteristics.” *Genomics* 109 (3–4): 186–91. https://doi.org/10.1016/j.ygeno.2017.03.001.

Tian, Shulan, Huihuang Yan, Claudia Neuhauser, and Susan L. Slager. 2016. “An Analytical Workflow for Accurate Variant Discovery in Highly Divergent Regions.” *BMC Genomics* 17 (1). https://doi.org/10.1186/s12864-016-3045-z.

Torkamaneh, Davoud, Jérôme Laroche, Maxime Bastien, Amina Abed, and François Belzile. 2017. “Fast-GBS: A New Pipeline for the Efficient and Highly Accurate Calling of SNPs from Genotyping-by-Sequencing Data.” *BMC Bioinformatics* 18 (1): 5. https://doi.org/10.1186/s12859-016-1431-9.

Torkamaneh, Davoud, Jérôme Laroche, and François Belzile. 2016a. “Genome-Wide SNP Calling from Genotyping by Sequencing (GBS) Data: A Comparison of Seven Pipelines and Two Sequencing Technologies.” *PLOS ONE* 11 (8): e0161333. https://doi.org/10.1371/journal.pone.0161333.

Treangen, Todd J., and Steven L. Salzberg. 2012. “Repetitive DNA and Next-Generation Sequencing: Computational Challenges and Solutions.” *Nature Reviews Genetics* 13 (1): 36–46. https://doi.org/10.1038/nrg3117.

United States Department of Agriculture. 2019. “FY 2019 Budget Summary.” https://www.obpa.usda.gov/budsum/fy19budsum.pdf

Van der Auwera, Geraldine A., Mauricio O. Carneiro, Christopher Hartl, Ryan Poplin, Guillermo del Angel, Ami Levy-Moonshine, Tadeusz Jordan, et al. 2013. “From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline: The Genome Analysis Toolkit Best Practices Pipeline.” In *Current Protocols in Bioinformatics* 43:11.10.1-11.10.33. https://doi.org/10.1002/0471250953.bi1110s43.

Varala, Kranthi, Kankshita Swaminathan, Ying Li, and Matthew E. Hudson. 2011. “Rapid Genotyping of Soybean Cultivars Using High Throughput Sequencing.” *PLoS ONE* 6 (9): e24811. https://doi.org/10.1371/journal.pone.0024811.

Walling, Jason G., Randy Shoemaker, Nevin Young, Joann Mudge, and Scott Jackson. 2006. “Chromosome-Level Homeology in Paleopolyploid Soybean (*Glycine Max*) Revealed Through Integration of Genetic and Chromosome Maps.” *Genetics* 172 (3): 1893–1900. https://doi.org/10.1534/genetics.105.051466.

Wang, Chengzhong, Ramsey Najm, Qin Xu, Dah Eun Jeong, David Walker, Maureen E. Balestra, Seo Yeon Yoon, et al. 2018. “Gain of Toxic Apolipoprotein E4 Effects in Human IPSC-Derived Neurons Is Ameliorated by a Small-Molecule Structure Corrector Article.” *Nature Medicine* 24 (5): 647–57. https://doi.org/10.1038/s41591-018-0004-z.

Wang, K., M. Li, and H. Hakonarson. 2010. “ANNOVAR: Functional Annotation of Genetic Variants from High-Throughput Sequencing Data.” *Nucleic Acids Research* 38 (16): e164–e164. https://doi.org/10.1093/nar/gkq603.

Wickland, Daniel P., Gopal Battu, Karen A. Hudson, Brian W. Diers, and Matthew E. Hudson. 2017. “A Comparison of Genotyping-by-Sequencing Analysis Methods on Low-Coverage Crop Datasets Shows Advantages of a New Workflow, GB-EaSy.” *BMC Bioinformatics* 18 (1): 586. https://doi.org/10.1186/s12859-017-2000-6.

Wickland, Daniel P. and Yoshie Hanzawa. 2015. “The FLOWERING LOCUS T/TERMINAL FLOWER 1 Gene Family: Functional Evolution and Molecular Mechanisms.” *Molecular Plant* 8 (7): 983–97. https://doi.org/10.1016/j.molp.2015.01.007.

Wolfe, Kenneth H. 2001. “Yesterday’s Polyploids and the Mystery of Diploidization.” *Nature Reviews Genetics* 2 (5): 333–41. https://doi.org/10.1038/35072009.

Wu, Yongsheng, Felix San Vicente, Kaijian Huang, Thanda Dhliwayo, Denise E. Costich, Kassa Semagn, Nair Sudha, et al. 2016. “Molecular Characterization of CIMMYT Maize Inbred Lines with Genotyping-by-Sequencing SNPs.” *Theoretical and Applied Genetics* 129 (4): 753–65. https://doi.org/10.1007/s00122-016-2664-8.

Yang, Jian, Jian Zeng, Michael E. Goddard, Naomi R. Wray, and Peter M. Visscher. 2017. “Concepts, Estimation and Interpretation of SNP-Based Heritability.” *Nature Genetics* 49 (9): 1304–10. https://doi.org/10.1038/ng.3941.

Yang, Rui, Lee H. Chen, Landon J. Hansen, Austin B. Carpenter, Casey J. Moure, Heng Liu, Christopher J. Pirozzi, et al. 2017. “*Cic* Loss Promotes Gliomagenesis via Aberrant Neural Stem Cell Proliferation and Differentiation.” *Cancer Research* 77 (22): 6097–6108. https://doi.org/10.1158/0008-5472.CAN-17-1018.

Yu, Chien Hung, Yunkun Dang, Zhipeng Zhou, Cheng Wu, Fangzhou Zhao, Matthew S. Sachs, and Yi Liu. 2015. “Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-Translational Protein Folding.” *Molecular Cell* 59 (5): 744–54. https://doi.org/10.1016/j.molcel.2015.07.018.

Zhou, T, M Weems, and C O Wilke. 2009. “Translationally Optimal Codons Associate with Structurally Sensitive Sites in Proteins.” *Molecular Biology and Evolution* 26 (7): 1571–80. https://doi.org/10.1093/molbev/msp070.